

# Bayesian Approaches to Uncertainty in Speech Processing

Philip Neil Garner

PhD by Publication

University of East Anglia  
School of Computing Sciences



September, 2011

© 2011 by Philip N. Garner

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that no quotation from the thesis, nor any information derived therefrom, may be published without the author's prior written consent.

# Abstract

Many techniques in speech processing require inference based on observations that are often noisy, incomplete or scarce. In such situations, it is necessary to draw on statistical techniques that themselves must be robust to the nature of the observations. The Bayesian method is a school of thought within statistics that provides such a robust framework for handling “difficult” data. In particular, it provides means to handle situations where data are scarce or even missing.

Three broad situations are outlined in which the Bayesian technique is helpful to solve the associated problems. The analysis covers eight publications that appeared between 1996 and 2011.

Dialogue act recognition is the inference of dialogue acts or moves from words spoken in a conversation. A technique is presented based on counting words. It is formulated to be robust to scarce words, and extended such that only discriminative words need be considered.

A method of incorporating formant measurements into a hidden Markov model for automatic speech recognition is then outlined. In this case, the Bayesian method leads to a re-interpretation of the formant confidence as the variance of a probability density function describing the location of a formant.

Finally, the Gaussian model of speech in noise is examined leading to improved methods for voice activity detection and for noise robustness.

# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Background . . . . .	7
1.2	The Bayesian method . . . . .	8
1.2.1	Historical note . . . . .	8
1.2.2	Bayesian vs. classical . . . . .	8
1.2.3	A basic problem . . . . .	9
1.2.4	The merit of priors . . . . .	10
1.3	Applications of Bayesian methods . . . . .	12
1.3.1	Bulk mail . . . . .	12
1.3.2	Neural networks . . . . .	13
1.3.3	ASR . . . . .	13
1.3.4	Summary . . . . .	14
1.4	Motivation for the thesis . . . . .	14
<b>2</b>	<b>Dialogue</b>	<b>17</b>
2.1	Introduction . . . . .	17
2.2	Relevance . . . . .	17
2.3	Overlap . . . . .	18
2.4	Paper walk-through . . . . .	18
2.5	Analysis . . . . .	21
2.5.1	Method . . . . .	21
2.5.2	The small sample problem . . . . .	21
2.5.3	Dirichlet solution . . . . .	22
2.5.4	Vocabulary size . . . . .	23
2.5.5	The Poisson solution . . . . .	23
2.5.6	Zipf prior . . . . .	23
2.5.7	Language modelling and vocabulary pruning . . . . .	24
2.5.8	Improved prior distributions . . . . .	25
2.5.9	Summary . . . . .	25
2.6	With hindsight . . . . .	25
2.6.1	Multinomial . . . . .	25
2.6.2	Priors . . . . .	26
2.6.3	Language modelling . . . . .	26
2.7	Impact . . . . .	26
<b>3</b>	<b>Formant analysis</b>	<b>29</b>
3.1	Introduction . . . . .	29
3.2	Relevance . . . . .	29
3.3	Overlap . . . . .	30
3.4	Paper walk-through . . . . .	30
3.5	Analysis . . . . .	32
3.5.1	The Holmes formant analyser . . . . .	32
3.5.2	The confidence problem . . . . .	32
3.5.3	Convolving normal distributions . . . . .	33
3.5.4	Training . . . . .	34

3.5.5	Summary . . . . .	34
3.6	With hindsight . . . . .	34
3.7	Impact . . . . .	35
<b>4</b>	<b>Noise robustness</b>	<b>37</b>
4.1	Introduction . . . . .	37
4.2	Relevance . . . . .	37
4.3	Overlap . . . . .	38
4.4	Paper walk-through . . . . .	38
4.5	Analysis . . . . .	41
4.5.1	The Gaussian model of speech in noise . . . . .	41
4.5.2	The Sohn-Sung VAD . . . . .	42
4.5.3	VAD Implementation . . . . .	42
4.5.4	Differential VAD . . . . .	44
4.5.5	Evaluation metric . . . . .	44
4.5.6	From VAD to noise robustness . . . . .	44
4.5.7	Initial noise robustness work . . . . .	45
4.5.8	Cepstral normalisation . . . . .	46
4.5.9	Features and databases . . . . .	47
4.5.10	Summary . . . . .	48
4.6	With hindsight . . . . .	48
4.6.1	VAD . . . . .	48
4.6.2	SNR features . . . . .	48
4.6.3	Capacity of a Gaussian channel . . . . .	48
4.7	Impact . . . . .	49
<b>5</b>	<b>Conclusions</b>	<b>51</b>
5.1	Hypotheses . . . . .	51
5.2	Corollaries . . . . .	52
<b>A</b>	<b>Full list of publications</b>	<b>53</b>
<b>B</b>	<b>Letters from co-authors</b>	<b>57</b>
<b>C</b>	<b>Papers</b>	<b>62</b>

# Acknowledgements

I keep this brief.

The work in this thesis was done over the course of more than fifteen years at three establishments in three countries. Speech processing is a difficult topic requiring a specialised infrastructure to make any progress; regardless of the number of authors, all the work was collaborative. Numerous people helped and contributed; listing them would risk overlooking some, however many are indicated either as authors, or in the acknowledgement sections of the manuscripts. Others appear as co-authors on publications in the appendix.

It remains to acknowledge the people without whom this thesis would not have been written: Roger Moore, Wendy Holmes, Aidan Hemsworth and Toshiaki Fukada for being kind enough to provide letters acknowledging co-authorship. I am indebted to Hervé Bourlard for his support and encouragement at Idiap, and of course to Stephen Cox at UEA for his advice throughout.

Finally, I extend my gratitude to the examiners, Philip Jackson and Gavin Cawley, without whose comments the thesis would have been incomplete.

# Papers

The eight papers presented in this commentary fall into three clear categories.

## Dialogue move recognition

Philip N. Garner, Sue R. Browning, Roger K. Moore, and Martin J. Russell. A theory of word frequencies and its application to dialogue move recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1880–1883, October 1996.

Philip N. Garner and Aidan Hemsworth. A keyword selection strategy for dialogue move recognition and multi-class topic identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1997.

Philip N. Garner. On topic identification and dialogue move recognition. *Computer Speech and Language*, 11:275–306, 1997.

## Formant analysis

John N. Holmes, Wendy J. Holmes, and Philip N. Garner. Using formant frequencies in speech recognition. In *Proceedings of EUROSPEECH*, volume 4, pages 2083–2086, September 1997.

Philip N. Garner and Wendy J. Holmes. On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, 1998.

## Noise robustness

Philip N. Garner, Toshiaki Fukada, and Yasuhiro Komori. A differential spectral voice activity detector. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, May 2004.

Philip N. Garner. SNR features for automatic speech recognition. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, December 2009.

Philip N. Garner. Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition. *Speech Communication*, 53(8):991–1001, October 2011.

Of the above, all but the last two papers were written whilst employed in an industrial setting. This leads to bias towards conference papers rather than journal. The final two were written whilst at Idiap.

# Chapter 1

## Introduction

In this thesis, the use of Bayesian inference is investigated in three distinct applications in the field of speech technology, namely dialogue act recognition, formant analysis and noise robustness. The unifying theme common to each of these investigations is the characterisation of uncertainty, which not only allows more robust inference where only limited data are available, but also provides a unifying framework allowing techniques to fit together in a hierarchical manner. This chapter provides a brief review of the Bayesian methods used in the remainder of the thesis.

### 1.1 Background

In around 1993, I joined Andrew Webb's pattern processing applications group at what was DRA (the Defence Research Agency) in Malvern (it was formerly the Royal Signals and Radar Establishment, and subsequently the Defence Evaluation and Research Agency and QinetiQ). I worked on some basic pattern recognition techniques, and on a radar problem.

To address a (self-) perceived deficiency in statistics, I attended a one week course given by Anthony O'Hagan. O'Hagan taught Bayesian statistics of course, but he was also a proponent of a particular camp within the field known as subjective prior elicitation. This technique is quite pragmatic; rather than require purely data-driven methods, it draws on what might be described as anecdotal or heuristic evidence too. For instance, if you want to know the accuracy of a drill, the first thing you do is ask the drill operator how accurate he thinks it is.

Of course, O'Hagan covered other methods too, for instance weak priors and numerical integration methods. However, the subjective aspect fixed in my mind because it was at odds with, for instance, Andrew Webb, who was a maximum likelihood proponent, and David MacKay, whose evidence framework was gaining attention at the time.

The collection of papers presented here began after that course, and represent my putting into practice techniques that were taught there, and were also used within the pattern processing groups in Malvern. The three sections are rather disparate in application (except that they are all speech processing), but the underlying approach is the same.

## 1.2 The Bayesian method

### 1.2.1 Historical note

The term Bayesian takes its name from a paper attributed to the Reverend Thomas Bayes (Bayes, 1763), although Bayes's famous theorem does not appear at all. Rather, what Bayes did was discuss the concept of inverse probability. The theory was invented independently by the marquis de Laplace (1812), and is built around the following rather simple relationship: if there are two events,  $a$  and  $b$ , that are not independent, the joint probability of the two events can be written

$$P(a, b) = P(a | b) P(b) \quad (1.1)$$

or

$$P(a, b) = P(b | a) P(a). \quad (1.2)$$

Equating the two, we have:

$$\underbrace{P(a | b)}_{\text{Posterior}} = \frac{\overbrace{P(b | a)}^{\text{Likelihood}} \overbrace{P(a)}^{\text{Prior}}}{\underbrace{P(b)}_{\text{Evidence}}}, \quad (1.3)$$

where the terms are often referred to as indicated. That is, if we want to use knowledge of  $b$  to infer something about  $a$ , we can use knowledge of  $a$  to infer something about  $b$  in combination with priors (see below). There are a great many texts on Bayesian methods; for the purposes of this thesis I have mainly used O'Hagan (1994), but perhaps the most ubiquitous is that of Bernardo and Smith (2000).

### A note on terminology

In discussing Bayesian statistics, it is natural to refer to the theorem attributed to Thomas Bayes. I believe there are three possible ways to do this:

**Bayes's theorem** The theorem belongs to Bayes; cf. Newton's first law.

**Bayes' theorem** As above, but following a precedent used for, e.g., Greek names, which tend to end in  $s$ , where the final  $s$  is omitted.

**Bayes theorem** A named theorem; *the* Bayes Theorem. cf. the Poincaré conjecture.

I have used these interchangeably.

### 1.2.2 Bayesian vs. classical

Bayesian statistics is often contrasted with the frequentist (or classical) approach, where probability is defined as being the result of a large number of successive tests. In fact, Bayes's theorem has meaning in classical statistics too: It is the basis of a hypothesis test. As long as  $a$  and  $b$  are events, all is well. If  $a$  is a parameter,  $\theta$ , however, and  $b$  is data,  $d$ , difficulty occurs in the classical situation: Whilst the value  $p(d | \theta)$  can have meaning (we have several examples of  $d$ ), the value  $p(\theta | d)$  does not have meaning because we



cannot sample  $\theta$  directly. Thus Bayes's theorem does not apply to parameters in the classical approach; rather, enough examples of  $d$  are required.

Implicit in the Bayesian approach, however, is a definition of probability as being a degree of belief; just a number where 0 represents certainty that a proposition is false and 1 represents certainty that a proposition is true. In the Bayesian approach, there is nothing to prevent  $p(\theta | d)$  being expanded via Bayes's theorem. This leads to two fundamental features of the Bayesian method for parameters:

1. Inference proceeds from the posterior  $p(\theta | d)$  rather than the likelihood  $p(d | \theta)$ .
2. A prior,  $p(\theta)$ , can (in fact, must) be specified representing a degree of belief over different values of the parameter.

The key is that, whilst classical approaches distinguish parameters and data (or events), the Bayesian approach treats them equally. Bayes's theorem can applied to either.

One advantage of the Bayesian approach follows directly: Classical approaches are reliant upon  $p(d | \theta)$  being well defined by being based on *enough* samples. The Bayesian approach is able to augment this measurement using a prior, so when *enough* samples are not present a robust result can still follow.

### 1.2.3 A basic problem

The Bayesian method in the context of this thesis is well illustrated in the context of a problem where training and test data are related by an unknown parameter.

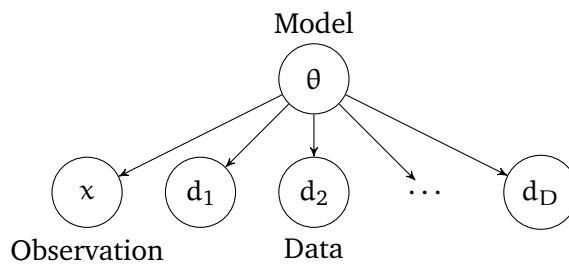


Figure 1.1: An inference diagram for a simple parametric model.

Say we have an observation,  $x$ , and we need inference based on  $x$ . Typically this means we are interested in  $p(x)$ , the probability density of  $x$ . With reference to figure 1.1,  $x$  was generated by a model with parameter  $\theta$ . If  $\theta$  is known, we can simply write down  $p(x)$  as the *conditional* density

$$p(x) = p(x | \theta), \tag{1.4}$$

and this is what the model generates. If  $\theta$  is unknown, however, it becomes a *nuisance variable* and it is necessary to *marginalise*:

$$p(x) = \int d\theta p(x | \theta) \underbrace{p(\theta)}_{\text{Prior}}. \tag{1.5}$$

As indicated in the equation, the marginalisation requires a *prior* on  $\theta$ . In this case, the prior is posterior to known training data  $\mathbf{d} = \{d_1, d_2, \dots, d_D\}$ . The situation is actually

better written

$$p(x | d_1, d_2, \dots, d_D) = \int d\theta p(x | \theta) p(\theta | d_1, d_2, \dots, d_D) \quad (1.6)$$

$$p(x | \mathbf{d}) = \int d\theta p(x | \theta) p(\theta | \mathbf{d}). \quad (1.7)$$

In order to evaluate the final term, we invoke Bayes's theorem:

$$p(x | \mathbf{d}) = \int d\theta p(x | \theta) \frac{p(\mathbf{d} | \theta) p(\theta)}{p(\mathbf{d})} \quad (1.8)$$

Equation 1.8 is sometimes called the Bayesian *predictive distribution*.

Three important concepts (at least for speech processing) follow from the predictive distribution:

**1. Point estimation** If  $p(\theta | \mathbf{d})$  can be assumed to be close to a delta function at  $\hat{\theta}$ , then equation 1.7 ceases to be an integral. Rather, it is just the likelihood given the point estimate of the parameter:

$$p(x | \mathbf{d}) = p(x | \hat{\theta}), \quad (1.9)$$

where we do not yet specify how the point estimate is obtained.

**2. MAP estimation** In the above case, a point estimate is required. The *maximum a-posteriori* (MAP) estimate is that which maximises the fractional part of equation 1.8:

$$\hat{\theta} = \operatorname{argmax}_{\theta} \frac{p(\mathbf{d} | \theta) p(\theta)}{p(\mathbf{d})}. \quad (1.10)$$

Notice that the denominator can be ignored.

**3. ML estimation** In the case that  $p(\mathbf{d} | \theta)$  can be assumed to be a delta function (i.e., at least compared to the prior), then we can write

$$\hat{\theta} = \operatorname{argmax}_{\theta} p(\mathbf{d} | \theta), \quad (1.11)$$

which is the *maximum likelihood* (ML) estimate of  $\theta$ .

So, the concept of ML training, which is a classical technique, drops out of a rigorous Bayesian perspective given a couple of assumptions. As a “bonus”, there is also MAP training that can be thought of as a Bayesian influenced ML approach. This also ties down the concept of *enough* samples from above. *Enough* is enough such that the delta functions are good approximations, and in the MAP approach, that concept of enough can come from the prior too.

#### 1.2.4 The merit of priors

To illustrate the effect of the prior, consider the task of finding whether or not a coin is biased after having seen the result of only one flip. Denote the probability of a head by  $\rho$ . In the general case of H heads and T tails, the likelihood is

$$p(H, T | \rho) = \rho^H (1 - \rho)^T. \quad (1.12)$$

This function has a maximum at

$$\hat{\rho} = \frac{H}{H + T}. \quad (1.13)$$

We are interested in the value of  $p(\rho | H, T)$ . Suppose the flip yielded a head, so  $H = 1$  and  $T = 0$ .

**1. ML estimation** In the classical ML framework, all that can be done is state that an estimate,  $\hat{\rho}$ , of  $\rho$  is the maximum of  $p(H, T | \rho)$ . So we have the value

$$\hat{\rho}_{ML} = 1 \quad (1.14)$$

directly from equation 1.13, implying complete bias towards heads.

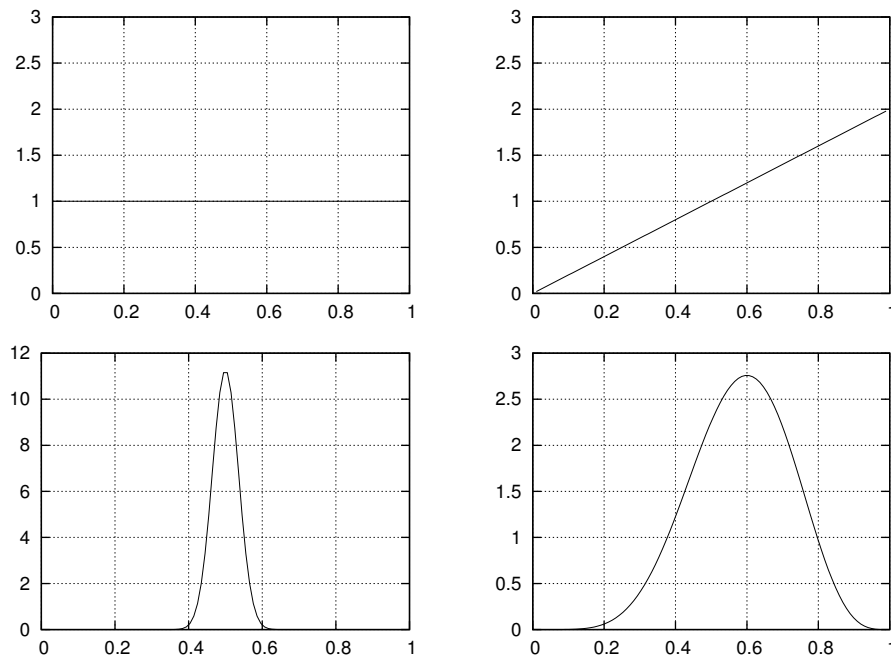


Figure 1.2: Beta distributions; in each case the abscissa is  $\rho$  and the ordinate is  $p(\rho)$ . Top left: Flat prior  $p(\rho) \propto 1$ . Top Right: Posterior  $p(\rho | H, T)$  for a flat prior and single flip yielding heads. Bottom left: Prior with  $\alpha = \beta = 100$ . Bottom right: Prior with  $\alpha = 7$  and  $\beta = 5$ .

**2. Flat prior** Now assume that we can proceed in a Bayesian sense, so

$$p(\rho | H, T) = \frac{1}{p(H, T)} p(H, T | \rho) p(\rho). \quad (1.15)$$

Setting  $p(\rho) \propto 1$  (figure 1.2, top left), the maximum of  $p(\rho | H, T)$  is still at 1, but it is now a function (figure 1.2, top right); there is finite probability that  $\rho$  can take any value between 0 and 1 (or more correctly,  $0 < \rho \leq 1$ ;  $\rho$  cannot be zero). This is important; it implies that the bias is probably not 1 at all. It is clear from the figure that there is a probability of 0.25 that  $\rho < 0.5$ .

**3. Informative unbiased prior** In fact,  $p(\rho) \propto 1$  does not represent prior knowledge at all; there is strong prior knowledge that the coin is unbiased. This can be represented by a

beta distribution

$$p(\rho) \propto \rho^{\alpha-1}(1-\rho)^{\beta-1}, \quad (1.16)$$

where  $\alpha = \beta$  and  $\alpha \gg 1$ . Figure 1.2, bottom left, shows such a distribution with  $\alpha = \beta = 100$ . Again following equation 1.15, we find that  $\rho$  can take any value between 0 and 1, but the function has a clear peak at

$$\hat{\rho}_{\text{MAP}} = \frac{H + \alpha - 1}{H + T + 2\alpha - 2} = \frac{\alpha}{2\alpha - 1}. \quad (1.17)$$

For large  $\alpha$ , this is very close to  $\hat{\rho}_{\text{MAP}} = 0.5$ , so the single flip does not change the prior very much.

**4. Informative biased prior** Say there is some prior knowledge that a coin is biased; it appears to be generating more heads than tails. This potentially alters the prior in two ways:

1. The maximum of the prior should be shifted away from 0.5 towards, say, 0.6. So  $\alpha \neq \beta$ .
2. The prior should be widened, representing uncertainty, so  $\alpha$  and  $\beta$  should be much smaller, closer to 1.

Say  $\alpha = 7$  and  $\beta = 5$  (which gives a peak at 0.6, illustrated in figure 1.2, bottom right). The posterior distribution now has a peak at

$$\hat{\rho}_{\text{MAP}} = \frac{H + 7 - 1}{H + T + 12 - 2} = \frac{7}{11} \quad (\approx 0.636). \quad (1.18)$$

There are two points here:

1. The initial “estimate” for a single flip is quite close to the prior.
2. Because the prior is wider, it would only take a few more flips (around 10) for the data to have a significant effect.

This “biased” prior, then has an effect somewhere between the “flat” and “unbiased” priors. More generally, it follows that an informative prior allows stronger conclusions to be drawn with fewer data, but only if the prior is correct. Conversely, an incorrect prior will require more data to reach the same strength of conclusion.

## 1.3 Applications of Bayesian methods

### 1.3.1 Bulk mail

Probably the most ubiquitous application of the Bayesian method in use today is the Bayesian “Spam Filter” of Sahami et al. (1998). In that work, a rule based system for classifying bulk email is replaced with a probabilistic one. In fact, the system is a combination of a naive Bayes classifier, attributed to Good (1965), and a vector space model. So, it is not Bayesian in the sense of attaching priors to parameters, but the rigorous framework is present. It is worth stressing that the term “naive” in this context refers to the features used by the classifier being assumed independent. The authors felt the need to justify the naive assumption,

but in practice it appears to be perfectly reasonable in the sense that the resulting system works.

### 1.3.2 Neural networks

An earlier (and more thorough) example of the Bayesian method is its application to neural networks by MacKay (1991, 2003), popularised by Bishop (1995). MacKay's multi-layer perceptron (MLP) was useful because it could produce error-bars on the outputs owing to its actually producing PDFs. However, his main contribution was perhaps more subtle: At the time, it was popular to use a technique known as weight decay to penalise large weights in the MLP. MacKay showed that weights could be penalised by means of a (Gaussian) prior. This led to a very similar training mechanism, but with the added security of knowing what the assumptions were. In this sense, he was showing that an ad-hoc technique could be made rigorous by means of the Bayesian framework.

### 1.3.3 ASR

Automatic speech recognition (ASR) is perhaps the most persuasive illustration of the power of both statistical rigour and the Bayesian approach. The first ASR systems were template based (Holmes and Holmes, 2001). That is, a set of templates of individual words were stored. A new word was compared with each template using a dynamic programming procedure, yielding a score representing a distance between the new word and each template. The new word was classified as being the same as the template corresponding to the lowest score.

The question then arose: How can this be generalised to the case where a whole sentence is spoken? In particular, how can the fact that some sentences are more common than others be accommodated. Whilst one can imagine various ad-hoc solutions, the key insight is that it can be trivially represented as a hypothesis test: We have a set,  $\mathcal{G}$ , of grammatical entries (words), and some observation sequence,  $\mathbf{a}$ . The task is choose the grammatical sequence,  $\mathbf{g}$ , selected from  $\mathcal{G}$ , with highest likelihood:

$$\hat{\mathbf{g}} = \operatorname{argmax}_{\mathbf{g}} p(\mathbf{g} | \mathbf{a}) = \operatorname{argmax}_{\mathbf{g}} \frac{p(\mathbf{a} | \mathbf{g}) p(\mathbf{g})}{p(\mathbf{a})}. \quad (1.19)$$

So, Bayes's theorem naturally leads to the concept of an acoustic model  $p(\mathbf{a} | \mathbf{g})$  and a language model  $p(\mathbf{g})$ . It is described by Bahl et al. (1983) amongst others. Further, the template and dynamic programming can be seen as the evaluation given a point estimate of section 1.2.3 and replaced by hidden Markov models (HMMs) and the Viterbi algorithm (Viterbi, 1967; Forney, 1973). The HMMs in turn are trained using the ML estimation of section 1.2.3 (Baum et al., 1970). The language model is normally N-gram based; Chen and Goodman (1996) give an older but thorough review.

Given the rigorous formulation, it is then possible to bring in other techniques that would otherwise be rather difficult to introduce. For instance, the concept of maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1994) is now ubiquitous in ASR, but would have been impossible without the statistical formulation.

Thus far, the above is statistically rigorous but not Bayesian. However, it has enabled Bayesian approaches, especially in adaptation. The first was the MAP approach of Gauvain

and Lee (1992, 1994), which takes a necessarily Bayesian interpretation of the acoustic model parameters, allowing use of small amounts of extra data to tune the HMM parameters towards a particular situation. Later, the linear transform used for MLLR adaptation was also interpreted in a Bayesian sense, allowing the transform to be trained on even smaller amounts of data. The current state of the art is probably the structural MAP approach of Siohan et al. (2002).

It is worth emphasising that HMMs are known not to be particularly good models of speech. Rather, their success stems from the fact that they can be trained easily. Further, that being more rigorous about their training leads to improved performance suggests that, at least in this case, the suitability of the model is secondary to treating the model properly.

### 1.3.4 Summary

The above implies a set of principles governing the Bayesian method. Certainly the model should be explicit, relating model parameters to data via a generative mechanism. However, it is not important that the model is accurate; rather the analysis should reflect the assumptions. Unknown variables should be removed by marginalisation; MAP and ML are approximations to this. Where inference follows the opposite direction to generation, Bayes's theorem should be used to "invert" probability. Finally, note that priors are necessary both for marginalisation and inverse probability.

## 1.4 Motivation for the thesis

Taking the applications described above together, it is possible to discern a generalisation as follows:

1. A simple problem can be solved using an ad-hoc but intuitive formulation. The ad-hoc formulation does not easily support more complicated cases.
2. Reformulating with statistical rigour allows more complicated cases to be supported. Other techniques that are themselves statistically rigorous fit into the framework easily.
3. Taking a Bayesian approach to parameters allows further benefits in terms of data sparsity.

This leads to two hypotheses<sup>1</sup> for the thesis:

1. Where an existing technique is somehow ad-hoc or not rigorous, we hypothesise that making it rigorous will lead to benefit in terms of allowing extensions that would not be possible otherwise.
2. Where an existing technique is rigorous, but not Bayesian, we hypothesise that making it Bayesian will lead to benefit in terms of robustness to small sample sizes.

Each of the papers in the following chapters, whilst targeting its own end, goes some way to investigating these hypotheses. That is, each paper is written from a stance that the hypotheses are true. Whilst the hypotheses are extremely general, they do provide a means to bind the papers in the thesis, and to draw conclusions.

---

<sup>1</sup>Of course, these hypotheses are being stated after the work has been done. Nevertheless, I believe they are representative of an approach taken at the outset of each piece of work in the thesis.

## Bibliography

- Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, March 1983.
- Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.
- Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763. Communicated by Richard Price in a letter to John Canton.
- José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons Ltd., 2000.
- Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.
- Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modelling. In *Proceedings of the 34th annual meeting of the ACL*, June 1996.
- G. David Forney. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, March 1973.
- Jean-Luc Gauvain and Chin-Hui Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11:205–213, June 1992.
- Jean-Luc Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov models. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, April 1994.
- I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, 1965.
- John Holmes and Wendy J. Holmes. *Speech Synthesis and Recognition*. Taylor & Francis, 29 West 35th Street, New York, NY 10001, 2nd edition, 2001.
- C. J. Leggetter and P. C. Woodland. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG/TR. 181, Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, England, June 1994.
- David J. C. MacKay. *Bayesian Methods for Adaptive Models*. PhD thesis, California Institute of Technology, 1991.
- David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- Pierre-Simon marquis de Laplace. *Théorie analytique des probabilités*. (Available from Dover), 1812.
- Anthony O’Hagan. *Bayesian Inference*, volume 2B of *Kendall’s Advanced Theory of Statistics*. Edward Arnold, 1994.

- Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A Bayesian approach to filtering junk e-mail. In *AAAI'98 Workshop on Learning for Text Categorization*, 1998.
- Olivier Siohan, Tor André Myrvoll, and Chin-Hui Lee. Structural maximum a posteriori linear regression for fast HMM adaptation. *Computer Speech and Language*, 16(1):5–24, January 2002.
- Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, IT-13(2):260–269, April 1967.



# Chapter 2

## Dialogue

### 2.1 Introduction

Around 1995, I began to work closely with the speech research unit (SRU) in Malvern. In fact, the pattern processing groups and SRU were very closely related, the former having grown out of the latter. The project was called spoken language understanding and dialogue, which had the (rather unappealing) acronym SLUD.

SLUD was about investigating techniques that might be useful in data-driven processing of events at a higher semantic level than phones or words. It grew from the observation that automatic speech recognition (ASR) tended to be data-driven and worked well. By contrast, language processing tended to be rather hand crafted and worked less well. Further, attempts at hand crafting ASR tended not to help. It was better to design algorithms that could learn from data.

At the time, the SRU had had success in topic spotting, and in a technique that involved using dynamic programming (DP) to find long sequences that were discriminative of topic (Nowell and Moore, 1993, 1995). The original plan was to use these techniques to distinguish dialogue moves (more often called dialogue acts). The moves were those in the dialogue games of Kowtko et al. (1993), using the HCRC (Human Communication Research Centre) Map Task Corpus (Anderson et al., 1991). Initial results had been published by Bird et al. (1995).

Our initial approach was to take a step back from the DP sequences, and find a baseline using some very simple feature. The language modelling concepts of unigrams and bigrams were good candidates. At the time, It was possible to submit incomplete papers to ICASSP (the International Conference on Speech and Signal Processing); we submitted one with blank tables of results headed 1-gram, 2-gram, DP features etc. It was rejected; this turned out to be a good thing, because we did not progress beyond unigrams.

### 2.2 Relevance

In the theme of this analysis, the dialogue move recognition problem is a classic statistical small sample size problem. There are observations (words) from which inference is required, but no data is available. Given the lack of training data, techniques based on parameter estimation simply do not work; parameters must be marginalised.

There are two key points from a Bayesian point of view:

**Choice of model** Being explicit about the generative model leads to the correct inference solution. It is possible to change the model slightly (from multinomial to Poisson), in which case the statistical approach changes.

**Choice of prior** In this case, the prior is simple but necessary. It is an objective prior (data-driven).

## 2.3 Overlap

There is some overlap between the papers in this section. In particular:

1. The core work presented in Garner et al. (1996) formed the basis of the project, and was later extended into the journal submission (Garner, 1997), so many of the results are repeated.
2. Many of the dialogue related results from Garner and Hemsworth (1997) are repeated in Garner (1997).

However, Garner and Hemsworth (1997) contains an analysis of the LOB (Lancaster Oslo Bergen) corpus, and results using absolute discounting that do not appear elsewhere. Garner (1997) includes new results based on a log-linear prior, and new results for vocabulary pruning based on equal move probability.

It might be argued that Garner et al. (1996) is redundant given Garner (1997). I include it because

1. It represents a chronological and appropriate approach to the research, testing material at a conference before submitting to an archival journal and evolving the idea.
2. It is shorter and hence easier to read.
3. It has other authors, reflecting the fact that, whilst it is mainly my own work, it was not done in isolation.

With regard to the final point, and with the benefit of experience, I believe it was a mistake not to include the other authors on the final journal submission.

## 2.4 Paper walk-through

### **1996: A theory of word frequencies and its application to dialogue move recognition**

Philip N. Garner, Sue R. Browning, Roger K. Moore, and Martin J. Russell. A theory of word frequencies and its application to dialogue move recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1880–1883, October 1996.

As the title suggests, we thought this represented a new theory of word counting; the application to dialogue moves was secondary at the time. The prose proceeds extremely quickly, almost skipping the background and motivation.

The paper begins with a very quick overview of the application area; that is, an interpretation of the HCRC map-task corpus as a classification task. The map-task corpus is annotated with dialogue moves; we wanted to infer the dialogue move given the words. The section 2.2 on methodology simply states that the maximum likelihood classification task is dictated by Bayes's theorem, and that the model is based on sequentially sampling a categorical random variable. Results are presented immediately (at the top of page 2!) and the discussion points out that something is wrong. In particular, too many utterances are classified as 'Ready', and this is counter-intuitive.

Section 3 proceeds to add some rigour to what is thus far somewhat intuitive. In particular there are three insights in section 3:

1. In calculating probabilities in the intuitive way, one is actually assuming a particular model — a multinomial.
2. If that is the assumption, there is a right way to express that assumption mathematically. Whilst it is still intuitive, it is not obvious.
3. The unknown vocabulary,  $V$ , can be removed by making the same approximation that relates the Poisson and binomial distributions, yielding a multiple-Poisson.
4. Using this slightly different assumption, there is a right way to approach the mathematics.

Section 4 presents a rather more detailed description of Zipf's law as a means of incorporating prior knowledge in the formulation. Zipf's law states that a rank ordering of word frequencies is roughly reciprocal square root in shape. A figure is presented showing that Zipf's law broadly holds over a several unrelated databases.

It is shown that Zipf's law can be reinterpreted to represent a prior for the multiple Poisson distribution, but it cannot quite be represented by the (conjugate) gamma distribution.

In section 5, two evaluations are presented. The first is a repeat of the evaluation of section 2 showing that the problems associated with the intuitive solution are addressed by the Poisson distribution, i.e., there is no counter-intuitive favour for the class 'Ready'. The second experiment equalises the amount of data per class (dialogue act), then plots performance against amount of training data. This shows that both the reformulation and the prior have quite significant benefit.

### **1997a: A keyword selection strategy for dialogue move recognition and multi-class topic identification**

Philip N. Garner and Aidan Hemsworth. A keyword selection strategy for dialogue move recognition and multi-class topic identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1997.

The paper opens by discussing topic identification, which is by definition a two-class problem (the classes are 'Wanted' and 'Unwanted'). The introduction also states that the two class problem can be solved by maximising a measure known as 'Usefulness', which is a likelihood ratio. This likelihood ratio is also a means to choose which keywords to use (hence the name). It is then stated that in the general  $M$ -class case, the decision rule backs off to Bayes's theorem, but there is no general definition of usefulness to determine the

choice of keywords. The remainder of the paper is about how to define usefulness for the M-class problem.

Section 2 introduces some information-theoretic measures that intuitively appear capable of functioning as usefulness:

1. Mutual information can represent the information provided about a dialogue move event by a word event.
2. Change in entropy is the increase in entropy of the ensemble given a word event.
3. Saliency is a weighted mutual information used by Gorin.

Section 3 then derives a new measure being the expected change in class posterior probability when a word is observed. This is intuitively reasonable, but also specifically maximises the decision rule. It is shown that this strategy when applied to a multinomial based measure yields the same usefulness measure that would be used intuitively. In the case of the Poisson distribution the form is tractable, but contains digamma functions.

Section 5 presents experiments. The basic technique was to gradually increase the size of the vocabulary using the given usefulness metrics, and evaluate classification accuracy vs. vocabulary. Two databases were used; the HCRC map task corpus, with classes being dialogue moves, and the LOB corpus with classes being report topics. For the multinomial tests, two options were available to handle out of vocabulary words (OOVs); a default count of 0.5, or absolute discounting (from the language modelling literature). For the Poisson case, OOVs were taken care of via the gamma prior. Results favoured the Poisson distribution, showing that the generalisation of usefulness extended to that case. Subsequent discussion focusses on the comparatively poor performance of mutual information, and the fact that change in entropy can work well.

## **1997b: On topic identification and dialogue move recognition**

Philip N. Garner. On topic identification and dialogue move recognition. *Computer Speech and Language*, 11:275–306, 1997.

This paper in *Computer Speech and Language* is effectively a consolidation of the above two papers; however, all sections are expanded.

The first few sections follow Garner et al. (1996), but adding context and discussion. This places the work with respect to other literature, and puts the theory on a sound footing. The two class usefulness is explained from first principles, showing it to be a loss minimisation strategy.

In section 3, the generative model associated with the multinomial is explained. An experiment is presented for the naive (or intuitive) case, and the unknown word problem is discussed in section 3.3. The section concludes by showing that the Bayesian (or maximum a-posteriori, MAP) solution to the multinomial (which is a dice-throwing model) is dependent upon the vocabulary.

Section 4 introduces the multiple Poisson in detail, explaining the derivation from the multinomial; section 5 then discusses Zipf's law and how to use it to represent a prior. Two possible prior forms are discussed; the first is a gamma distribution that is conjugate to the Poisson likelihood. The second is an ad-hoc log-linear prior that fits the data better, but is not conjugate. Word sequence probabilities are given for each of the two possible priors.

Section 6 presents essentially the same evaluation that was first presented by Garner et al. (1996), except that this time the log-linear prior is included. The log-linear prior is shown not to have any benefit over the conjugate prior.

The final technical section, 7, presents a summary and extension of the work presented in Garner and Hemsforth (1997). The multi-class discriminability measure is derived in detail for both the multinomial and Poisson distributions. Figure 6 is a summary of the original results; the multinomial and Poisson on the same graph emphasising the difference. Figure 7 is new; it shows the same information, but for equal move probability. Figure 8, however, is taken directly from the conference paper.

The conclusions state that the Poisson based measures of both probability and discriminability provide a better foundation than ad-hoc or multinomial ones. Further, numerical results from other authors are placed into context.

## 2.5 Analysis

### 2.5.1 Method

When I began working on SLUD, the other project members had thought about the dialogue problem and about the features, but less so about the discrimination method. In particular, they wanted to view the problem as an expanded topic spotting problem. This was itself problematic because topic spotting is fundamentally a two class problem; dialogue move recognition is multi-class. Further, the two class nature of topic spotting leads naturally to likelihood ratio approaches, and this was the basis of the measure known as *usefulness* that was favoured at the time. Usefulness was useful because it was both a discrimination metric and an indicator of the utility<sup>1</sup> of a given feature. However, it was not clear how to apply usefulness in a multi-class problem.

In coming from a pattern recognition group, I was in a position to first sort out the discrimination method. This amounted to going back to first principles to derive usefulness, showing that it was just an application of maximum likelihood, and writing down the multi-class solution. It is section 2.2 (Methodology<sup>2</sup>) of Garner et al. (1996). Whilst it seems trivial now, at the time it was well received because it gave a sound basis to an area that had previously not been well understood.

### 2.5.2 The small sample problem

In doing dialogue move (or topic) discrimination using word frequencies, it is natural to try to attach a probability to each word in the sentence. The probability of the sentence is then the product of the probabilities of the component words. These component word probabilities come from training data. The naive approach is to count the number of times a given word appeared in data of a given move (call it  $n$ ), then divide by the total number of words in that move (call it  $N$ ), to give

$$P(w | n, N) = \frac{n}{N}, \quad (2.1)$$

---

<sup>1</sup>I think utility is probably a better word than usefulness, but it has a distinct meaning in information theory

<sup>2</sup>Which should really be just method; methodology is study of methods

where I am using  $P(w | n, N)$  here to loosely represent probability of word  $w$ . The notation in the papers is more rigorous.

The difficulty arises when one tries to do this for several moves. It is illustrated by the three “reply” moves:

**R-Y** is reply affirmative; it is a reply to a question and has the meaning *yes*. In practice, it is composed of words such as *yeah, yep, OK* and *all right*.

**R-N** is the opposite; it is a negative reply to a question. It tends to be composed of words like *no, nope* and *different*.

**R-W** is a more general reply to a question of the form *what/where/how* etc. It has a very general vocabulary; much like normal speech.

Given a common word such as *yes*, it is straightforward to apply a probability to this for moves R-Y and R-W. It is likely, however, that *yes* was never observed in the move R-N ( $n = 0$ ). This gives it a probability of zero, hence the whole sentence has probability zero. Clearly something is wrong; no word can have a probability of zero; small certainly, but not zero. The reason is data-sparsity; there simply isn’t enough training data to cover all words in all moves. Of course, the problem is well known in language modelling, although it typically occurs with higher order  $n$ -grams rather than unigrams.

As in the method case, the initial work had identified that there was a problem with small sample sizes, but had applied only a temporary fix. The fix was to assume that unobserved samples had actually occurred 0.5 times. This turned out to have come from a conversation with someone with a statistics background who had understood the topic spotting application, understood the counting problem, and given a quick solution based on a working knowledge of statistical regularisation techniques. It was never meant to be an authoritative answer.

### 2.5.3 Dirichlet solution

In looking at the sentences in a Bayesian manner, it is natural to look for a generative model for the words. A good initial model is a dice-throwing model, which is represented by a multinomial distribution. At any given time, each word has a given probability of being selected. The probabilities of all words sum to unity.

It turns out that if one assumes a multinomial and derives the ML solution, one ends up with the same  $n/N$  solution that the naive approach gives. This is important though; it indicates that in dividing  $n$  by  $N$ , one is implicitly assuming a multinomial model. However, the Bayesian approach also points to what is missing: With no contribution from the likelihood term, the inference comes from the prior.

The common conjugate prior for a multinomial is a Dirichlet distribution. It turns out to be straight-forward to represent a flat prior using a Dirichlet distribution, in which case the maximum a-posteriori (MAP) estimate becomes

$$P(w | n, N) = \frac{n + 1}{N + V}, \quad (2.2)$$

where  $V$  is the vocabulary size. Notice that when  $n = 0$ , the probability is not non-zero. However, there is now a new parameter,  $V$ . This is explained in section 3.1 of Garner et al. (1996).

#### 2.5.4 Vocabulary size

The revelation that the vocabulary size was important prompted a rather long and fruitless search for a means to find it.

The language modellers' approach is to fix  $V$  to some convenient number, then place all out of vocabulary (OOV) words into an unknown word class. This works for automatic speech recognition (ASR) because the vocabulary is assumed closed. In dialogue move recognition, however, the vocabulary is open. Rather, we need the vocabulary from which the words are being drawn; this is the total number of words known to all the speakers.

It turned out that various work had been done by statisticians in counting species. That problem can be stated as follows: A biologist is collecting data at the edge of an inaccessible forest. In one day, he has seen (say) 20 birds, 3 small mammals and one large mammal. How many species are there in the forest? In our case, the species are words and the number of species is the vocabulary; otherwise it is exactly the same problem. These techniques had been applied by Efron and Thisted (1976) specifically for finding the vocabulary of William Shakespeare.

Although the species counting techniques were interesting, one conclusion was that the resulting vocabularies simply depended on initial assumptions. It was not an exact science.

We also tried (unsuccessfully) to put a prior on  $V$ . The same problems arise, however: Which prior distribution, and with which parameters? Thankfully, a more elegant solution came to light.

#### 2.5.5 The Poisson solution

I had come across the way a Poisson distribution could be derived from a binomial. Although the reason to do this was originally to simplify computation, it was clear that another key point was that a binomial distribution has two parameters whereas a Poisson has only one. It struck me that the same thing could be applied to the multinomial.

The result, as detailed in Garner et al. (1996) section 3.2, is the multiple Poisson distribution. Its (crucial) advantage is that it can model known words whilst basically ignoring OOVs. Whilst the multiple Poisson was known as a distribution, it was certainly new in language processing. The Poisson distribution is usually used as an approximation to the binomial. In this case, however, it represented a change in the underlying assumption of the generative model.

#### 2.5.6 Zipf prior

Whilst the multiple Poisson neatly got around the unknown vocabulary problem, the question of the prior still remained. Indeed, the Poisson has a variable that can cover the whole positive real axis rather than just 0 to 1, so a flat prior is difficult to assign. The solution was to make use of Zipf's law. Although it tends to be used in language processing, Zipf's law is really just a power law that occurs in other fields too: It says that classes chosen at random are likely to be unlikely.

In fact, Zipf's law is about words arranged in rank order, but figure 1 in Garner et al. (1996) shows how to re-arrange the plot to get a probability density function (PDF) of word frequency corresponding to the Poisson parameter. We measured the slope empirically by using data from the internet, and fitting a conjugate gamma distribution by eye (figure 2 of

the same paper). Although the data-sets were quite small with hindsight (< 1 m words), data from quite different sources were found to correspond quite well.

Overall, the combination of the multiple Poisson and Zipf based prior worked well. Aside from leading to an improved overall move classification, perhaps more importantly, it also distributed the classifications more evenly across the available moves, evident in tables 1 and 2 of Garner et al. (1996).

### 2.5.7 Language modelling and vocabulary pruning

After publishing the first paper, it became clear that two issues were outstanding:

1. What we were actually doing was language modelling, albeit with a very simple model. It was necessary to evaluate standard language model techniques such as discounting.
2. We had been using the whole vocabulary of the task to perform classification. It had always been clear, however, that only discriminative words were necessary. What was less clear was how to find the discriminative words. In the two class multinomial case, the usefulness measure also indicated discriminative words, so it had never been an issue. The difficulty was in how to do it over multiple classes.

Both of these problems were approached by Garner and Hemsworth (1997), perhaps emphasising the latter.

Given the opportunity to supervise a new colleague, Aidan Hemsworth, I described this problem to him, and he took up the challenge (amongst other things). For the language modelling, we looked at discounting and simply implemented a standard technique. In addition, Aidan was interested in information theoretic measures for vocabulary selection, and came up with two possible ones: Mutual information and change in entropy. He was able to show that they could be used to select reasonable keywords. In assigning the information measure to each word, it seemed reasonable to average or take an expectation over the different classes. In any case, Aidan was able to build up a reasonable approach to the problem. We also compared with a related measure known as salience introduced by Gorin (1995).

For my part, I felt that it ought to be possible to choose words in a probabilistic sense rather than an information theoretic sense. For instance, the usefulness measure used in the two-class case, even though it *looked* like an information theoretic measure, was actually derived probabilistically.

Driven by the fact that usefulness suggested that the useful words were the ones that contributed most to the score, I tried an approach based on differentiating the probability measure. One key to this was to minimise the reciprocal of probability rather than maximise the probability itself. This is useful because Bayes's theorem has a product in the numerator and a sum in the denominator; the reciprocal separates into simpler terms.

Applying this method to the multinomial form of the problem resulted in virtually the same usefulness measure that was used in topic spotting (Garner and Hemsworth, 1997, section 3). This meant we had a generic method of finding the utility of a given word independently of the number of classes or underlying distributional assumptions. We applied the method to the Poisson based measure and, whilst the functional form was not trivial to compute, it was reasonably simple.



The performance of the new usefulness and information theoretic measures is illustrated in figures 1 and 3 of Garner and Hemsworth (1997). In summary, the combination of the Poisson based measure and multi-class usefulness perform far better than any other combination.

### 2.5.8 Improved prior distributions

The work thus far was consolidated into a journal article (Garner, 1997) that also included a section on priors.

Previously, I had approximated the plot from Zipf's law using a gamma distribution. This bothered me because it really wasn't right. Fundamentally, a gamma distribution rises polynomially and falls exponentially; by contrast, the Zipf plot falls polynomially. The gamma distribution can be made to fall polynomially, but only with a gradient greater than -1. This is illustrated in figure 3 of Garner (1997). Also illustrated is a better fitting line based on an ad-hoc but pragmatic distribution: A log-linear fit.

The log-linear prior was not conjugate; that is, it did not have a convenient functional form. This led to difficulties in doing the marginalisation. It is well known that Bayesians expend much effort, often with a copy of "Gradshteyn" (Gradshteyn and Ryzhik, 2000), searching for solutions to integrals. This was certainly the case here. However, the integral did turn out to have a solution in terms of confluent hypergeometric functions, and these can be evaluated using published algorithms.

The log-linear prior gave no *significant* improvement over the gamma prior; rather, on the whole it was worse. Certainly when taking into account the computational complexity it brought no extra value.

### 2.5.9 Summary

The three papers Garner et al. (1996), Garner and Hemsworth (1997) and Garner (1997) describe work in dialogue move recognition. The contributions are as follows:

**A multiple Poisson based model** for word frequencies that is independent of the vocabulary.

**A generalised measure of utility** for words that is independent of the underlying distribution and of the number of classes.

**A Zipf-based prior** that represents Zipf's law in terms of PDF of frequency, plus interpretations as gamma and log-linear distributions.

## 2.6 With hindsight

### 2.6.1 Multinomial

In some sense, I don't think I dealt fairly with the multinomial distribution. For instance, the Zipf based prior, suitably normalised for a probability, could have been applied to the multinomial to give a word probability similar to

$$P(w | n, N, V) = \frac{n + 0.1}{N + 0.1V}, \quad (2.3)$$

where the 0.1 is the gradient of the Zipf plot, or perhaps

$$P(w | n, N, V) = \frac{n + 1/V}{N + 1}, \quad (2.4)$$

with  $V$  chosen to be the vocabulary of the training set.

Certainly the multiple Poisson was neater, but I was also perhaps too keen to work on my own invention rather than apply the insights to the simpler model.

## 2.6.2 Priors

Given a distribution, it is possible to work out a weak (Jeffries) prior. The weak prior is usually quite simple, often improper. Typically for a positive quantity  $x$  such as a variance, the weak prior is the reciprocal of the quantity,  $x^{-1}$ . This is polynomial, not so different from the  $x^{-0.9}$  that comes from setting  $\alpha = 0.1$  in the gamma prior.

It is possible that Zipf's law is actually just a weak prior. Proving this would be an achievement in itself as it would tie together two natural but unrelated occurrences.

Another possibility would have been to use an inverse-gamma distribution. The inverse-gamma naturally rises exponentially and falls polynomially. It is, however, not conjugate; it would probably lead to a fairly complicated marginal distribution.

## 2.6.3 Language modelling

I did try the multiple Poisson in a language modelling problem. At the time, I could not get it to work better than a standard smoothed unigram model. However, I now know that ASR performance is dependent upon assumption of an open or closed vocabulary in the language modelling. It is possible that I simply made a mistake.

## 2.7 Impact

Google scholar is aware of 16 citations of the journal article; slightly fewer for each of the two conference papers.

I am aware of the following works that have been directly influenced by one or more of the three papers:

- The work was continued to an extent by Simon Smith and Martin Russell at the University of Birmingham, and is reported in the Ph.D. thesis of Smith (2003).
- The papers were used as the basis for several patents filed by Canon Inc. (Garner et al., 2000a,b,c).
- It was acknowledged by Allen Gorin in the context of his "How may I help you?" work at AT&T (Gorin et al., 1997).
- Probably the most high profile citation is by Bellegarda (2000) in Proceedings of the IEEE.

## Bibliography

- Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. The HCRC map task corpus. *Language and Speech*, 34(4):351–366, October/December 1991.
- Jerome R. Bellegarda. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296, 2000. Invited paper.
- Stuart Bird, Sue R. Browning, Roger K. Moore, and Martin J. Russell. Dialogue move recognition using topic spotting techniques. In *Proceedings ESCA Workshop on Spoken Dialogue Systems*, pages 45–48, Vigsø, Denmark, May 1995.
- Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, December 1976.
- Philip N. Garner. On topic identification and dialogue move recognition. *Computer Speech and Language*, 11:275–306, 1997.
- Philip N. Garner and Aidan Hemsworth. A keyword selection strategy for dialogue move recognition and multi-class topic identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1997.
- Philip N. Garner, Sue R. Browning, Roger K. Moore, and Martin J. Russell. A theory of word frequencies and its application to dialogue move recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1880–1883, October 1996.
- Philip Neil Garner, Jason Peter Andrew Charlesworth, and Asako Higuchi. Language recognition using sequence frequency. United States Patent 6882970, Canon Kabushiki Kaisha, October 2000a. URL <http://www.google.com/patents?id=GoUTAAAAEBAJ>. Issued: Apr 19, 2005.
- Philip Neil Garner, Jason Peter Andrew Charlesworth, and Asako Higuchi. Pattern matching method and apparatus. United States Patent 7212968, Canon Kabushiki Kaisha, October 2000b. URL [http://www.google.com/patents?id=1dd\\_AAAAEBAJ](http://www.google.com/patents?id=1dd_AAAAEBAJ). Issued: May 1, 2007.
- Philip Neil Garner, Jason Peter Andrew Charlesworth, and Asako Higuchi. Language recognition using a similarity measure. United States Patent 7310600, Canon Kabushiki Kaisha, October 2000c. URL <http://www.google.com/patents?id=2PifAAAAEBAJ>. Issued: December 18, 2007.
- A. L. Gorin, G. Riccardi, and J. H. Wright. How may i help you? *Speech Communication*, 23(1–2):113–127, October 1997.
- Allen L. Gorin. On automated language acquisition. *Journal of the Acoustical Society of America*, 97(6):3441–3461, June 1995.
- I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series and Products*. Academic Press, sixth edition, 2000. Alan Jeffrey, Editor.

Jacqueline C. Kowtko, Stephen D. Isard, and Gwyneth M. Doherty. Conversational games within dialogue. Technical report, Human Communication Research Centre, Edinburgh and Human Communication Research Centre, Glasgow, 1993.

P. Nowell and R. K. Moore. A non-word based approach to topic spotting in speech. Memorandum 4815, DRA, October 1993.

Peter Nowell and Roger K. Moore. The application of dynamic programming techniques to non-word based topic spotting. In *Proceedings Eurospeech'95*, volume 2, pages 1355–1358, Madrid, Spain, September 1995.

Simon G. J. Smith. *Predicting query types by prosodic analysis*. PhD thesis, The University of Birmingham, School of Electronic, Electrical and Computer Engineering, August 2003. URL <http://postgrad.eee.bham.ac.uk/smithsgj/Smith2003.pdf>.

## Chapter 3

# Formant analysis

### 3.1 Introduction

At the time I was working on the dialogue move recognition, another colleague, Wendy Holmes, was working on a segmental HMM. The segmental HMM was an attempt to get around the fact that a speaker independent HMM could generate (when sampled) consecutive frames from different people; clearly not physically realistic. The segmental HMM first defined a segment to be the time spent in a given state, so it was a semi-Markov model. Then for each segment it generated a sample from a speaker independent distribution; this sample was used as the mean of a (narrower) speaker dependent distribution. The effect was that within any segment, the samples were more representative of a single speaker rather than many speakers.

That the segmental HMM was a better model of speech led to the notion of using it for speech synthesis. It was (and is) in principle possible to build a codec comprising a speech recogniser followed by a speech synthesiser. This in turn required a feature extraction technique that could be inverted; at the time, MFCC (mel-frequency cepstral coefficient) based front-ends were not designed to do that.

In looking for a suitable feature extractor, Wendy had come across a formant analyser written by her father, John Holmes. Formants are an attractive feature for speech synthesis. They can be re-synthesised using a formant synthesiser. They also represent tangible and intuitively meaningful features, at least for speech scientists.

### 3.2 Relevance

In the context of this analysis, this chapter is an example of subjective prior elicitation. That is, the prior is a mathematical representation of the opinion of an informed user. This is related to the case of the drill operator of O'Hagan (1994): To find the accuracy of a machine, you can get a good initial estimate by asking the human operator. The difference is that, in this case, the human is interpreting the output of the machine in a probabilistic sense.

It is subjective; a different operator would interpret the output differently.

Modelling and marginalisation are also important: An explicit model dictates a correct approach, replacing an ad-hoc (albeit intuitively reasonable) one.

### 3.3 Overlap

Whilst the two papers are quite distinct in prose, the underlying technique is the same. However, there is very little technical overlap; they can be viewed as a single paper that describes the formant extractor followed by a mathematical exploration of how to incorporate the formant information into ASR.

### 3.4 Paper walk-through

#### 1997c: Using formant frequencies in speech recognition

John N. Holmes, Wendy J. Holmes, and Philip N. Garner. Using formant frequencies in speech recognition. In *Proceedings of EUROSPEECH*, volume 4, pages 2083–2086, September 1997.

This paper is an overview of the formant analyser in which the three authors each describe their respective work. My own contribution is in the second half (from section 3). The introduction proceeds by introducing formants, and stating that formants cannot work alone to distinguish certain speech sounds. However, for some other sounds (with a peaky structure), formants can clearly be of use. It goes on to say that formant tracking has pitfalls, in particular where formants are either not well defined, or not well distinguished. In such cases, however, continuity constrains ought to help.

Section 2 argues that human experts can usually label formants quite accurately up to a certain point. The limits are illustrated in figures 1, 2 and 3; in particular, the last figure shows that there is ambiguity when fewer formants are visible. In this latter case, the human expert can easily make several hypotheses.

The section continues (in 2.2) by describing an ad-hoc algorithm that labels formants in a similar manner as might be used by a human expert. The labelling is based on log power spectral frames of 64 point discrete Fourier transforms (DFTs), and the sequence proceeds as follows:

- Each frame is compared with around 150 hand labelled templates, the “few” closest ones are retained.
- These few are further compared using a dynamic programming (DP) approach. The template with the best DP score is retained.
- The frequency warping from the DP is applied to the frame at a 125 Hz quantisation.
- A finer quantisation is obtained by comparing with templates of formant shapes.

In 2.3, a further DP process is explained. Instead of running in the frequency dimension, this one runs in the time dimension and is concerned with formant continuity.

Section 3 introduces the concept of confidence estimates. During periods of silence, background noise, or no obvious spectral peaks, there is no confidence in the formant estimates. By contrast, for peaky spectra there is high confidence. This uncertainty can be represented as the variance of a notional Gaussian distribution of underlying frequency about the estimated value. It is explained that in the formant estimator, low confidence values are used to favour (ad-hoc) prior values for formants rather than measured ones. In the

recogniser, however, the variance associated with confidence is added to that of the model. This in turn is an improvement over a previous ad-hoc method.

In section 4, experiments are presented. The key point of the experiments is to use 8 cepstra, but replace the upper 3 cepstra with the 3 formant estimates; the lower 5 cepstra represent general spectral shape. Table 1 shows a logical progression of experiments where the performance of 5 cepstra is improved by adding 3 more cepstra, but can be improved much more (3 times more than for cepstra) by adding 3 formants instead. The confidence framework must be used, otherwise the results deteriorate by the addition of formants. Alternating hypotheses also help a little.

The discussion and conclusions essentially state that the formants behave as hoped, with the variation in error rate for various conditions being expected.

### **1998: On the robust incorporation of formant features into hidden Markov models for automatic speech recognition**

Philip N. Garner and Wendy J. Holmes. On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, 1998.

By contrast to the first paper, the second one was written largely by me, and is largely mathematical. It opens with an introduction of formants, and a short description of the potential shortfalls described in Holmes et al. (1997). It then makes a short case for the confidence measure having the potential to overcome these shortfalls. Section 2 goes on to state that the confidence measures are actually standard deviations, and these can be squared and used as variances of normal distributions centred on the formant estimates. High confidence represents small variance, low confidence is large variance.

Section 3 comprises the majority of the paper, describing the mathematical formulation in some detail. In 3.1 it is argued that the formant extractor viewed as a generative model emits both formant estimates and confidence estimates. This can equivalently be considered as a noisy channel model, where a formant is corrupted by zero mean Gaussian noise. The solution is shown to be a convolution of the model distribution with that implied by the confidence estimate. This leads to the confidence variance adding to the model variance.

Section 3.2 considers the re-estimation case. Given observations with confidences, how does one train an HMM set. The derivation follows that of Liporace (1982), first showing that the re-estimation of the transition probabilities is unchanged from the standard case. The expression for the model means is shown to depend upon the re-estimate of the variances; it is stated that the current value rather than the re-estimate can be used. In the case of the model variances, however, there is no obvious solution, and an approximation is necessary.

Two approximations are considered. Method 1 (section 3.2.1) makes the assumption that the confidence is time independent for a given state. Although this yields a solution, it is not necessarily positive definite. As the degenerate case is where the confidence is low, it is suggested that accumulation only happen for high confidence values. Method 2 (section 3.2.2) is simply a mathematical trick that allows the variance to be isolated, but also involves the same value being replaced by its current value elsewhere in the expression.

Section 4 briefly points out the corollary that method 1 ceases to be an approximation for the case of constant confidence. This is identical to having additive stationary Gaussian noise.

Experiments are presented in section 5 in the same manner as for Holmes et al. (1997): Five cepstra that describe general spectral shape are augmented by other cepstra, or by formants with various confidence handling. As reported in the previous paper, the confidence measure is necessary to allow formants to work at all, and when used they are beneficial. When the confidence is used in training too, a further benefit is observed, especially when second choice formants are included. Both variance re-estimation approximations work equally well.

## 3.5 Analysis

### 3.5.1 The Holmes formant analyser

The “usual” approach to formant extraction is to use, for instance, linear prediction, then project the pole positions onto the unit circle by solving a polynomial. This kind of approach has two drawbacks, however:

1. Formants are often not easily distinguished; in the extreme case, two formants can merge into a single peak. Linear prediction cannot distinguish them in this case.
2. Formants can often disappear; the often cited case is for nasals, where one formant simply ceases to exist.

These and similar cases are illustrated on page 2 of Holmes et al. (1997).

John Holmes had been interested in formants for many years, and was no doubt frustrated with automatic formant extraction techniques. I recall him telling me that, even in the cases above, he was able to look at a spectrogram and write down perfectly accurate formant positions. His answer to this frustration was to design a formant analyser that he could train in a very hands-on manner. It is described by Holmes and Holmes (1996), and subsequently by Holmes et al. (1997).

The Holmes analyser contained many (of the order of a hundred) templates in the form of hand annotated spectra. The templates were then warped in the same manner as dynamic time warping, but along the frequency axis, in order to match spectra under test. Templates with small warping scores, i.e., those that were not far away for the spectra under test, were then used to label the formant positions. Using this paradigm, John was able to simply add more templates when mistakes were made. Over time, a good reference template set had presumably been accumulated.

One side effect of the warping was that any formant estimate could also have associated with it a confidence value. The confidence was largely heuristic, but could be represented as a four bit number — an integer from 0 to 15. One extreme of confidence represented high confidence, being clearly defined formants; the other extreme was low (or zero) confidence, being ill-defined or even missing formants.

### 3.5.2 The confidence problem

In trying to use the formant analyser in the context of ASR, we came across the difficulty of how to incorporate the confidence measures. They were clearly important; they indicated



that some numbers should not be relied upon. How to use them was far from clear, however. The first attempt was to weight (in a multiplicative sense) the probability densities of the formant features by a numerical confidence value; high confidence was 1, low was 0. This had the effect of significantly reducing the likelihood of low confidence formants but, whilst it worked to an extent, was flawed in the sense that it was ad-hoc.

The solution was not immediately apparent, and I do not recall how I came upon it, but the answer appeared to be to interpret the confidence as a variance. In fact, the reciprocal of variance known as *precision* was probably a better analogy, but we went with variance because it was more familiar. The interpretation is as follows: The formant estimates are means or modes of normal distributions. The confidence values are variances. In cases of high confidence, the variances are very small (high precision), and the normal distributions approach delta functions. For low confidence, the variances become high (low precision). The high variance indicates that the formant can lie in quite a wide range. In the extreme case, the variance is infinite<sup>1</sup>, yielding a flat distribution; this is indicative of a formant simply not existing.

The interpretation is fundamentally Bayesian; it reinterprets the formant analyser as a degree of belief indicator. The actual underlying formant positions are unknown state variables; the formant tracker reports a degree of belief about these variables.

One key component of this interpretation was to attach variances to the confidence values. Both John and Wendy understood the concept, and were able to do this. The paper was a nice collaboration between the disciplines of speech signal processing and statistics. Although I did not have a deep understanding of the formant analyser, I was able to enhance it by providing a more rigorous statistical approach.

### 3.5.3 Convolution of normal distributions

Given the interpretation of the formant analyser and confidence values, it still remained to formulate how to use resulting distributions in an HMM based ASR system. In fact, the formulation developed in parallel with the interpretation. It is described in detail by Garner and Holmes (1998), and reduces to a fairly simple convolution.

If the model (HMM state) has mean  $m_m$  and variance  $v_m$ , and the observation (formant) is a mean  $m_o$  and a variance  $v_o$ , representing an unobserved  $x$ , the combination is a convolution of the two distributions:

$$\int_0^{\infty} dx \underbrace{\frac{1}{\sqrt{2\pi v_m}} \exp\left(-\frac{(x - m_m)^2}{2v_m}\right)}_{\text{Model distribution}} \underbrace{\frac{1}{\sqrt{2\pi v_o}} \exp\left(-\frac{(x - m_o)^2}{2v_o}\right)}_{\text{Observation distribution}} = \underbrace{\frac{1}{\sqrt{2\pi(v_m + v_o)}} \exp\left(-\frac{(m_m - m_o)^2}{2(v_m + v_o)}\right)}_{\text{Combined distribution with variances added}}. \quad (3.1)$$

That is, the variance from the confidence adds to the variance in the HMM state distribution. It is appealingly simple.

Further, the interpretation is clear: For high confidence, the formant variance is small, yielding the usual HMM calculation. For low confidence, however, the added variance is high, tending to flatten out (hence equate) all state distributions associated with that obser-

<sup>1</sup>In fact, I don't recall if the low confidence case was mapped quite to infinity.

vation. In turn, competing models receive the same probability density contribution from low confidence formants, but different contribution from high confidence ones.

Results were presented by Holmes et al. (1997) showing not only that the method worked, but that it allowed formant features to out-perform otherwise comparable cepstral features.

### 3.5.4 Training

Although we had proved that the variance method worked for ASR, it was in recognition only. The models had been trained using an ad-hoc method; I do not recall exactly how, but most likely by simply ignoring the confidence values. It was natural to try to incorporate the variance method in training. The formulation turned out to be simple enough, but no closed form solution existed for the mean and variance reestimation. In searching for a solution, I was able to postulate two approximations. Wendy (who was the one familiar with the code base) tried both approximations. The results are detailed by Garner and Holmes (1998). We found that incorporating the variance into training was beneficial, but not so much so as for recognition. This made sense; the training process was able to integrate errors over many frames rendering errors in any given frame insignificant.

One aside is that whilst formulating the training, I found a (minor) mistake in the work of Liporace (1982), which is seen as a kind of tutorial for reestimation by some people.

### 3.5.5 Summary

The two papers describe work in formant analysis. Overall there are perhaps three contributions:

**A formant model** that models formant estimates along with variances in those estimates.

**A recognition paradigm** where the formant variances are added to the (HMM) model variances to incorporate uncertainty.

**A training paradigm** where approximations are presented to incorporate the same uncertainty into model training.

## 3.6 With hindsight

It is a source of regret that this work was not consolidated into a journal article; it could have expanded and deepened the original papers as follows:

1. The formant method was only evaluated on two corpora of digits. Whilst this is a sensible start point, digits only have limited phonetic coverage. In turn, formants model some phones better than others.
2. The mapping between formant analyser internal features, confidence values and variance values could have been analysed more thoroughly. In particular, it may have been possible to derive variance values directly from the dynamic programming process in the spectral matching.
3. The influence of delta features could have been investigated. In particular, formants have much better defined dynamic properties than cepstra.

### 3.7 Impact

Formant tracking has never been a mainstream technique, but the attraction of such features still draws in researchers. When they publish, they do tend to cite this work; actually the first paper more than the second. Owing to this, these papers are cited more than any others with which I have been involved. Google scholar reports 47 and 34 citations for the two papers respectively.

In particular, I am aware of the following works that have been directly influenced by the two papers:

- Wendy Holmes went on to use the formant tracker in her work on segmental HMMs (Holmes, 2004).
- Nick Wilkinson and Martin Russell considered the problem of phone recognition in TIMIT (Wilkinson and Russell, 2002). In fact, they report poor results that could be improved by using the confidence as a weighting factor against MFCC features.
- Martin Russell and Philip Jackson at the University of Birmingham also used the formant tracker for segmental HMMs (Russell and Jackson, 2005).
- Ljubomir Josifovsky was aware of the work and cited it in his Ph.D. thesis (Josifovsky, 2002).
- Katrin Weber at Idiap cited it in the context of her HMM2 work (Weber et al., 2003).
- Perhaps most significantly, the papers are cited in the context of uncertainty decoding (Stouten et al., 2006). In uncertainty decoding, a variance is associated with noisy observations, usually in the context of the vector Taylor series (VTS) noise robustness technique. Given the variance, it is natural to use the same technique, although it was most likely rediscovered as the formulation is fairly well known.

The use in uncertainty decoding suggests that the training formulation may be useful for training in noise. In Garner and Holmes (1998), I noted that for constant noise variance one of the formulations ceases to be an approximation. This means that if noise can be assumed to lead to a constant variance in the cepstral domain then the formula can be used.

### Bibliography

Philip N. Garner and Wendy J. Holmes. On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, 1998.

John N. Holmes and Wendy J. Holmes. The use of formants as acoustic features for automatic speech recognition. In R. Lawrence, editor, *Proceedings of the Institute of Acoustics*, volume 18, pages 275–282, 1996. Part 9. Autumn Conference Speech & Hearing.

John N. Holmes, Wendy J. Holmes, and Philip N. Garner. Using formant frequencies in speech recognition. In *Proceedings of EUROSPEECH*, volume 4, pages 2083–2086, September 1997.

- Wendy J. Holmes. Segmental HMMs: Modeling dynamics and underlying structure in speech. In Mark Johnson, Sanjeev P. Khudanpur, Mari Ostendorf, and Roni Rosenfeld, editors, *Mathematical Foundations of Speech and Language Processing*, volume 138 of *IMA volumes in mathematics and its applications*, pages 135–156. Springer, 2004.
- Ljubomir Josifovsky. *Robust Automatic Speech Recognition with Missing and Unreliable Data*. PhD thesis, Department of Computer Science, University of Sheffield, UK, August 2002.
- Louis A. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory*, IT-28(5):729–734, September 1982.
- Anthony O’Hagan. *Bayesian Inference*, volume 2B of *Kendall’s Advanced Theory of Statistics*. Edward Arnold, 1994.
- Martin J. Russell and Philip J. B. Jackson. A multiple-level linear/linear segmental HMM with a formant-based intermediate layer. *Computer Speech and Language*, 19(2):205–225, April 2005.
- Veronique Stouten, Hugo Van Hamme, and Patrick Wambacq. Model-based feature enhancement with uncertainty decoding for noise robust ASR. *Speech Communication*, 48: 1502–1514, 2006.
- Katrin Weber, Shajith Ikbal, Samy Bengio, and Hervé Bourlard. Robust speech recognition and feature extraction using HMM2. *Computer Speech and Language*, 17(2–3):195–211, April–July 2003.
- N. J. Wilkinson and Martin J. Russell. Improved phone recognition on TIMIT using formant frequency data and confidence measures. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2121–2124, Denver, CO, USA, September 2002.

# Chapter 4

## Noise robustness

### 4.1 Introduction

After a brief period working for Canon in the UK, I moved close to the head office in Tokyo. At that time, the speech group was being “moved” from the research part of Canon into the development part. There was some pressure to commercialise recognisers that had been developed as research engines. In testing the recognisers, we came across the two major difficulties that befall all commercial ASR groups:

1. The recognisers had to be multilingual to work in all countries in which potential products were sold.
2. The products were aimed at potentially noisy environments and the recognisers were not very robust to noise.

The solution to the first problem was rather simple: Build acoustic models for all target markets. There was no research involved, just work.

The second problem was more involved; it turns out to be more like three interconnected problems:

**VAD** (voice activity detection) is to distinguish voice from simply background noise. It determines when someone is speaking, hence when a recogniser should be active.

**Noise estimation** is the estimation of the background noise that has to be distinguished from speech in VAD, or removed from the speech in noise removal.

**Noise removal** is the reduction of noise in corrupted speech to yield uncorrupted (or clean) speech.

Whilst at Canon, I tackled the first two of these moderately successfully. The third one I began at Canon, but was not able to get good results until well after moving to Idiap in 2007.

### 4.2 Relevance

In the context of this analysis, the noise robustness problem is a small sample size problem. In general, the task is to estimate a variance (of speech) given only one datum (observation frame).

In the first paper, A Bayesian framework is defined, but only a ML solution can be obtained. In the later papers, the Bayesian method is extended to marginalisation over the noise, and MAP estimation.

By contrast with the previous chapter, the approach to priors is either non-informative or objective. Priors can be calculated from data.

In some sense, the approach is inverted: Instead of assuming a model and following the right statistical method, the desire for a correct statistical method influences the design of the model (using signal to noise ratio rather than energy). This is similar to the dialogue approach in chapter 2, where the model was changed from multinomial to Poisson to enable a better solution.

### 4.3 Overlap

The first paper in this section (Garner et al., 2004b) is totally distinct, sharing only the Gaussian model, which is not a contribution. The second two have significant overlap, the second (Garner, 2011) being an archival journal of the first (Garner, 2009a). There are significant differences however: Garner (2009a) contains results using MFCCs and an analysis of a gamma based prior; Garner (2011) contains results using PLPs (perceptual linear prediction coefficients) on more databases (although it summarises the earlier MFCC based ones) and contains further discussion about the articulation index.

### 4.4 Paper walk-through

#### 2004: A differential spectral voice activity detector

Philip N. Garner, Toshiaki Fukada, and Yasuhiro Komori. A differential spectral voice activity detector. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, May 2004.

The paper is about voice activity detection (VAD), and opens with a discussion of the reasons for VAD. In mobile telephony it is to save bandwidth and in ASR it is to save computation. The spectral VAD of Sohn and Sung (1998) is then introduced as an appealing model-based VAD that has stood up well to various other standard ones. It is stated that there are two problems with the Sohn VAD: One is that it does not distinguish spectral shape; the other is that it assumes adjacent DFT bins are uncorrelated.

Section 2 goes on to describe the background theory. A decision theoretic framework involves a hypothesis,  $\mathcal{H}$ , corresponding to speech or non-speech; this enables a cost function to be formulated, and the expected cost to be minimised. The resulting function is similar to that of Sohn and Sung (1998). In 2.2, the Gaussian model is summarised, including the VAD decision rule. A parameter  $\kappa$  is introduced to handle correlation between adjacent DFT bins.

In section 3, the differential spectral modification is derived. This is based on a simple high-pass filter (HPF) running along each frame in the frequency dimension. It is argued that this will correct for the two problems described in the introduction. The derivation is shown to be a probabilistic change of variable; the required integral is stated to be too complicated, and an approximation is calculated by working with consecutive pairs of bins. The

derivation then proceeds as two special cases depending on whether the gradient is positive or negative, the resulting density being a double-sided exponential. The VAD likelihood ratio follows trivially.

The evaluation (section 4) is performed on a proprietary database of 3360 one word utterances each embedded in 5 second recordings. 6 noise conditions are presented, 3 of which are much noisier than the other 3. The evaluation is a pragmatic metric based on whether the identified region overlaps with the actual speech region in the signal. The VAD is implemented in both power spectral and mel spectral domains, and they are compared. The noise estimator is a modified version of that of Sohn and Sung (1998). Results indicate that in relatively clean conditions there is no gain over the usual Gaussian VAD. In noisier conditions, however, the differential formulation is beneficial. There is no evidence to prefer the mel spectral domain over power spectral domain.

### **2009a: SNR features for automatic speech recognition**

Philip N. Garner. SNR features for automatic speech recognition. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, December 2009.

The paper is quite long for a conference paper at 6 pages. It begins with a discussion of noise reduction in the context of ASR. Cepstral mean normalisation (CMN) is discussed for convolutional noise, and it is stated that it also works well in additive noise, especially with variance normalisation (CVN, or CMVN together). The Gaussian model and spectral subtraction (SS) are discussed for additive noise. Noise tracking and histogram normalisation are also discussed.

Section 2 presents a simplistic analysis of CMN in the context of additive noise. It is shown that the feature presented to the decoder in this context is actually  $\log(1 + \text{SNR})$ . It is then suggested that since this is the case, it makes sense to calculate signal to noise ratio (SNR) at the outset, rather than rely on CMN to do it.

Section 3 then presents a more rigorous analysis. The Gaussian noise model is introduced, and it is shown that SS results as a maximum likelihood (ML) solution for the speech variance. This suggests that it is the variance that is required (or is sufficient) for the decoder. Motivated by this, the SNR is defined to be the ratio of variances of speech and noise and a similar derivation is done for SNR. In sections 3C and 3D, an alternative derivation of basically the same expression is shown to result from marginalising over the noise, an operation not possible for SS.

Section 3E shows that it is also possible to associate a prior distribution with the SNR, and a maximum a posteriori (MAP) solution is derived using a gamma density as the prior. Although it requires solution of a cubic, an analytical solution is possible; a means to set the hyperparameters is also detailed. The gamma prior is shown to discourage higher SNR values.

Experiments are presented in section 4 on the Aurora-2 database<sup>1</sup>. Aurora-2 is a noisy version of the TIDIGITS data; it was used to evaluate contenders for the ETSI advanced DSR<sup>2</sup> front-end (ETSI, 2002). The database and MFCC based front-end are described, together with four techniques under test:

---

<sup>1</sup><http://aurora.hsnr.de/>

<sup>2</sup>ETSI: European Telecommunications Standards Institute. DSR: Distributed Speech Recognition

1. MFCC baseline
2. Spectral subtraction
3.  $\log(1 + \text{SNR})$  with an ML solution for SNR
4.  $\log(1 + \text{SNR})$  with a MAP solution for SNR

Results are shown with CMN and with CMVN. Subsequent discussion (section 5) points out that SNR based features can outperform all other approaches. Use of a prior can further improve results, but it is doubtful whether the associated complexity is worthwhile. It is hypothesised that the benefit is coming from the relative placement of noise compensation and filter-bank.

Before the conclusion, the noise tracker is discussed in section 5B. It is stated empirically that a correction factor for the noise tracker is cancelled by the flooring of the noise, and this leads to a solution without hyper-parameters. It is stated that no proof is evident.

In conclusion, SNR features appear to work well in noisy conditions when combined with CMVN.

## **2011: Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition**

Philip N. Garner. Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition. *Speech Communication*, 53(8):991–1001, October 2011.

This paper is an archival journal version of Garner (2009a). Section 1 opens by introducing both additive and convolutional noise reduction. It states that, from practical experience, it is very difficult to beat CMVN for noise robustness, even though there is no good reason for it to be useful for additive noise. The Gaussian model is briefly introduced along with the goals of the paper.

Section 2 proceeds to essentially repeat the background of Garner (2009a), except that a second, speaker dependent, convolutional noise is distinguished from the usual channel noise. This serves to show that, whilst SNR features can remove channel noise directly, there is a convolutional noise that must still be left to CMN; hence SNR features are not a substitute for CMN. Section 3 goes on to largely repeat the formal derivation of SNR features from Garner (2009a).

In section 4, the paper diverges from the conference version into a discussion of the context of the SNR features. First, it is explained that the SNR is exactly the a-priori SNR that is ubiquitous in speech enhancement. The approach of Lathoud et al. (2005) is then discussed as a closely related, but more complicated model. Perhaps the most important discussion is the comparison with the articulation index (AI); it is pointed out that the two have the same form up to the linear transform. This in turn suggests a convergence between psycho-acoustics and known best practice in ASR.

The noise tracker is discussed at some length. It is shown that a minimum tracker, a noise estimator and the AI are all related via a constant modification of the SNR. Therefore, a heuristic optimisation of this parameter cannot be attributed to just one of these entities. Finally, it is stated that the interpretation of CVN in the context of SNR is trivial: It normalises for dynamic range.



In section 5, experiments are presented. Results from the conference paper using MFCC features on Aurora-2 are summarised. It is stated that previous experiments on Aurora-2 are unreliable because of the artificial and constrained nature of that database. Further, PLP features are quite common nowadays and merit investigation. Experiments are then presented based upon two clear hypotheses: to test beyond the limits of Aurora-2, and to test PLPs. The background of PLPs is discussed, and the feature extraction mechanism is presented.

First, results are presented for PLPs on Aurora-3, which is still digit based, but in real noise. Expected results a-priori from Aurora-2 are stated, and it is shown that performance on Aurora-3 fits broadly with the expectations, except for an improvement in otherwise matched conditions. Results on Aurora-4 are then presented, which by contrast is large vocabulary, but with artificially added noise. Again, a-priori expectations from Aurora-2 are borne out quite well. Overall, PLP features appear to benefit more from the SNR-cepstrum than do MFCC features. Finally, it is briefly stated that good results could not be obtained on meeting data with rich text evaluation. This is attributed to meeting data not being so noisy and train-test conditions being matched. This is again in line with predictions from Aurora-2.

In section 6, a few issues arising from the experiments are discussed. It is stated that results are not state of the art, but are intended only for comparison; the databases are standard and other results are in the literature.

The difference between taking SNR before and after the filter-bank is discussed. It is hypothesised that the SNR-spectrum lends itself to coloured noise, and the noises in the test sets are suitably coloured.

Finally, the PLP power law is discussed. It was found experimentally that not using a compression in PLP, usually cube root, is beneficial in noise. It is hypothesised that compression is not a noise robust technique, and in this scenario the noise is more important than perceptual issues.

In conclusion, SNR-spectral features have some advantages over spectral features, and lend themselves to a Bayesian analysis. They perform well in combination with CMVN in noisy conditions. The SNR-cepstrum can be seen as a form of AI, and uses features known in enhancement too.

## 4.5 Analysis

### 4.5.1 The Gaussian model of speech in noise

The core statistical model is worth emphasising as it is the basis for the whole of this noise robustness section. Assume that the input to a DFT is white Gaussian noise. As any linear combination of Gaussian variates is also Gaussian, each output of the DFT is Gaussian. Now relax the model a little to say that the DFT bins are still Gaussian, but their variances can differ (which would not be the case for input white noise). The model is now capable of representing coloured noise.

Assume that a DFT operation produces a vector,  $\mathbf{x}$ , with complex components,  $x_1, x_2, \dots, x_F$ , where the real and imaginary parts of each  $x_f$  are i.i.d. normally distributed with zero mean

and variance  $v_f$ . That is,

$$p(x_f | v_f) = \frac{1}{\pi v_f} \exp\left(-\frac{|x_f|^2}{v_f}\right). \quad (4.1)$$

In the case where two coloured noise signals are distinguished, a background noise,  $\mathbf{n}$ , and a signal of interest,  $\mathbf{s}$ , typically speech, denote the noise variance as  $\nu$  and the speech variance as  $\sigma$ . In general, the background noise can be observed in isolation and modelled as

$$p(n_f | \nu_f) = \frac{1}{\pi \nu_f} \exp\left(-\frac{|n_f|^2}{\nu_f}\right). \quad (4.2)$$

The speech, however, cannot normally be observed in isolation. It is always added to noise. When both speech and additive noise are present the variances add, meaning that the total signal,  $t_f = s_f + n_f$ , can be modelled as

$$p(t_f | \sigma_f, \nu_f) = \frac{1}{\pi(\sigma_f + \nu_f)} \exp\left(-\frac{|t_f|^2}{\sigma_f + \nu_f}\right). \quad (4.3)$$

#### 4.5.2 The Sohn-Sung VAD

The investigation of Garner et al. (2004b) began as an attempt to improve VAD. At the time, we used a VAD that simply thresholded energy, and was hence sensitive to noise levels. Increased noise meant increasing the threshold. In looking for a promising replacement we came across a VAD design by Sohn and Sung (1998). The Sohn VAD had two appealing properties:

1. It was noise adaptive; it contained a noise tracker that continually estimated background noise.
2. It was based on an explicit statistical model.

The Sohn VAD's core likelihood ratio, equation 8 in Garner et al. (2004b), is just the ratio of equations 4.3 and 4.2, accumulated across each frame:

$$L(t) = \prod_{f=1}^F \frac{\nu_f}{\sigma_f + \nu_f} \exp\left(\frac{\sigma_f}{\sigma_f + \nu_f} \cdot \frac{|t_f|^2}{\nu_f}\right), \quad (4.4)$$

where all terms are per frame.

#### 4.5.3 VAD Implementation

The Sohn VAD was actually introduced in two relatively short papers: Sohn and Sung (1998) introduced the basic concept, including an innovative noise tracker. Later, Sohn et al. (1999) added an HMM based hangover scheme, and a decision-directed speech power estimator. The implementation began as a superset of the all these techniques. None were so difficult to implement, and the published results suggested that all together would work well.

However, in implementing the Sohn VAD, several issues arose. These issues are not detailed in Garner et al. (2004b), but they serve to connect that paper with those of Sohn and colleagues.

The first is to do with the VAD metric requiring estimates of the variances  $\sigma$  and  $\nu$ . In the case of the noise, the estimate can come from periods of speech inactivity. For the speech

variance, however, the estimate favoured by Sohn et al. (1999) was the decision directed (DD) estimator of Ephraim and Malah (1984). The DD estimator is ubiquitous in speech enhancement; it is actually defined in terms of SNR, but in spirit it is of the form:

$$\hat{\sigma}_{f,i} = \rho \hat{\sigma}_{f,i-1} + (1 - \rho) (|t_{f,i}|^2 - \nu_{f,i}), \quad (4.5)$$

where  $t$  is the time frame and  $\rho \approx 0.98$ , i.e., it is an infinite impulse response (IIR) filtered spectral subtraction. We found that for VAD, no filter was necessary; just the SS of Sohn and Sung (1998) SS worked well, although it was necessary to floor it as with the usual ASR noise robustness calculation.

The VAD of Sohn et al. (1999) also made use of an HMM-based state machine. The machine had two states corresponding to speech and non-speech. However, only forward likelihoods (“alphas”) were calculated. The reverse pass that would normally calculate “gammas” was omitted. This had the effect of behaving like another IIR filter; there was a definite lag in the VAD. Of course, this had the effect that the authors required: It held the VAD on for a little while after the speech appeared to have stopped. However, it also caused a delay in the VAD firing in the first place. We found that a more conventional state machine of the type in figure 4.1 worked better. The basic function is that unless the

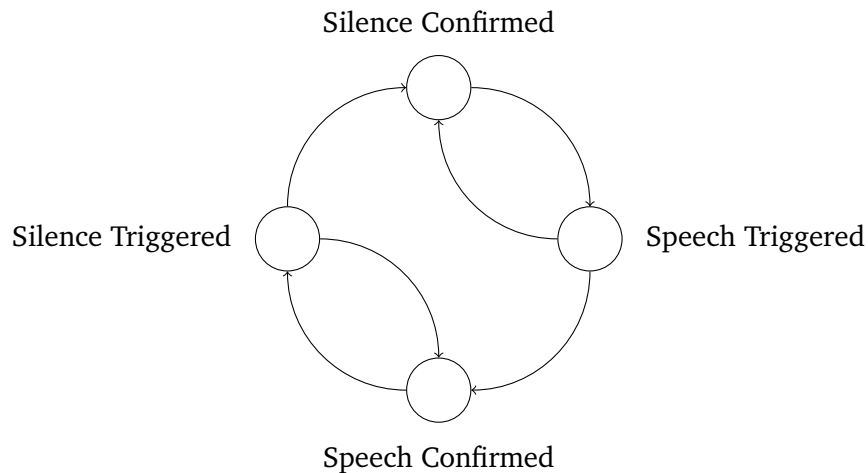


Figure 4.1: VAD state machine

VAD metric remains valid for a minimum time, the machine will revert to the previous state rather than advance to a confirmed state.

So, after much experimentation, the implementation was much closer to the original one of Sohn and Sung (1998) than that of Sohn et al. (1999).

Another difficulty was with the noise tracker. Sohn and Sung (1998) derive a noise estimator as

$$\hat{\nu}_{f,i} = \frac{1}{1 + L(t)_i} |t_{f,i}|^2 + \frac{L(t)_i}{1 + L(t)_i} \hat{\nu}_{f,i-1}, \quad (4.6)$$

which is another IIR filter. Recall that  $L(t)$  can approach zero for pure noise, and can be very high for high SNR. So, in high SNR conditions the noise estimate remains static, but for noise there is a time constant that approaches zero. The solution to this was to introduce a floor of  $\rho_\nu$ :

$$\hat{\nu}_{f,i} = \frac{1 - \rho_\nu}{1 + L(t)_i} |t_{f,i}|^2 + \frac{\rho_\nu + L(t)_i}{1 + L(t)_i} \hat{\nu}_{f,i-1}, \quad (4.7)$$

so, when  $L(t)_i$  is small, the time constant is defined by  $\rho_v$ . Equation 4.7 appears in Garner et al. (2004b), but the above explanation does not.

#### 4.5.4 Differential VAD

The material in the previous section is not in the paper; we did not consider it novel enough to try to publish. However, a differential VAD modification was worth publishing, the background of which is described below.

In testing the VAD on a variety of noise conditions, we found that it was rather sensitive to gain. That is, it was necessary to set a somewhat arbitrary threshold, and this threshold depended on the type and level of background noise. My colleague, Toshiaki Fukada, had suggested high-pass filtering the power spectrum as a solution to this. In turn, I had come across a paper by Nadeu et al. (2001) where the authors had described filtering along the frequency axis of each frame. In fact, filtering is an additive process rather than multiplicative, so it would not correct for gain as such, but it did seem to present advantages.

The initial difficulty with the filter was that it appeared to be a heuristic on top of what was otherwise a rather nice probabilistic model. However, that difficulty could be turned into an advantage by realising that the filter was actually a probabilistic change of variable. The filter is initially the difference between two power bins; equation 10 in Garner et al. (2004b). The joint probability is then equation 13 in the paper, and the change of variable is equation 16.

A further difficulty then arises: To apply the change of variable to the whole frame, the functional form becomes combinatorially intractable. This is at least in part because equation 16 in the paper has two alternate forms depending upon the sign of the filtered value. The solution was to consider only distinct pairs of power spectra; reducing the number of observations from, say, 128 to 64 (or 32 mel bins to 16).

#### 4.5.5 Evaluation metric

It is worth mentioning the evaluation metric. VAD is often measured in terms of ROC (receiver operating characteristic) curves, or RMS (root mean square) deviation of speech-silence boundaries. We used something we termed “gross error”, which was more pragmatic.

The general idea was that VAD *for ASR* can often be made to work by adding a suitably wide collar around the region indicated to be speech. This is especially true for short command and control utterances of the sort that we were interested in. It follows that as long as the VAD indicates part of the required speech region, it is doing well. Figure 1 in Garner et al. (2004b) is my formalisation of a set of conditions introduced by my colleague Toshiaki Fukada. They define acceptable and unacceptable performance conditions in terms of the start and end of speech detection. Whilst it was never proven or reported numerically, we found that gross error had a good correlation with word error rate.

#### 4.5.6 From VAD to noise robustness

The work on VAD was by no means independent of noise robustness, but used a rather unsophisticated noise robustness model. In tackling noise robustness in earnest, we were looking for a pragmatic solution. By this, I mean that there are two rather extreme approaches to noise robustness:

1. A simple approach, using techniques such as spectral subtraction or Wiener filtering.

Whilst many authors had reported success, our experience at Canon was that the performance was variable at best. We had an in-house test-set recorded in a few reasonable noise conditions. Techniques tended to work well for some noise conditions, but less well for others. SS and Wiener filters are characterised by ML solutions that need ad-hoc regularisation before they work at all.

2. A complicated approach, typified by the vector Taylor series (VTS) approach of Moreno et al. (1996).

If results in the literature are to be believed, this approach works very well indeed. However, VTS techniques are characterised by requiring a large Gaussian mixture acting as a prior on the speech signal. At Canon, we were trying to make the recognisers smaller; such a large mixture was prohibitive.

Somewhere between these two approaches was the likes of the ETSI advanced DSR front-end (ETSI, 2002). The ETSI system worked well, but was over-complicated, lacked theoretical rigour and was encumbered by patents.

It seemed to me that there ought to be a simple solution to the noise robustness problem, at least for stationary white or coloured noise. This solution should involve one or both of the following:

1. There should be a prior on the speech variance. This ought to give a solution in the spirit of SS or Wiener filtering, but with the ad-hoc regularisation replaced by justifiable hyperparameters.
2. It should be possible to marginalise over the noise.

This also followed from something I knew was wrong with the VAD work: Although it was model based and used inverse probability, it did not marginalise over nuisance parameters, and did not use priors. Worse, it involved point estimates (of likelihood ratios) based on point estimates (of the speech variance) based on a single datum.

#### 4.5.7 Initial noise robustness work

There is a significant time period between the publication of Garner et al. (2004b) and Garner (2009a). During that time (amongst other things), I failed to advance the noise robustness work significantly. It is instructive to briefly describe something that does not work!

Consider the parameters of 4.3; ignoring the  $f$  subscript for simplicity, there are two unknowns:

1. A noise variance,  $\nu$ .
2. A speech variance,  $\sigma$ .

Certainly the noise variance is a nuisance variable and should be marginalised. The speech variance is then either a parameter to estimate, or to marginalise if a Wiener-like solution is

required. However, marginalisation leads to the following difficulty:

$$p(t | \sigma) = \int_0^{\infty} d\nu p(t | \sigma, \nu) p(\nu | A, B) \quad (4.8)$$

$$= \int_0^{\infty} d\nu \frac{1}{\pi(\sigma + \nu)} \exp\left(-\frac{|t|^2}{\sigma + \nu}\right) \frac{B^A}{\Gamma(A)} \nu^{-A-1} \exp\left(-\frac{B}{\nu}\right) \quad (4.9)$$

where  $p(\nu | A, B)$  is an inverse-gamma distribution that results from estimating the PDF of  $\nu$  as in equations 16 and 17 of Garner (2009a);  $A$  and  $B$  represent the noise sample statistics. The integral above evaluates to an incomplete gamma function, which in turn is very difficult to do anything else with. It is the common Bayesian intractable integral problem. The same sort of issue arises when placing an (inverse-) gamma prior on  $\sigma$ .

The closest I came to a solution was to assume an estimate of  $\nu$  was possible (because there are usually several noise frames), then form a MAP estimate of  $\sigma$ . That work is written up as Garner (2009b); it was never published because the results do not show any improvement over SS.

#### 4.5.8 Cepstral normalisation

In trying to find priors for the speech variance, one of the most frustrating things was that I could never beat cepstral variance normalisation (CVN) in terms of accuracy in noise. There was no good reason for CVN to work, except that it “made the observation fit the model”. Worse, the combination of whatever prior I used and CVN was worse than just CVN.

In researching the area, I came across a paper by Lathoud et al. (2005) in which good results were reported using a measure based on SNR. The paper is actually more about a mixture model for speech and noise; no insight is presented about the SNR. However, it reminded me of something I had noticed much earlier; it is described in section 2A of Garner (2009a):

$$\begin{aligned} \log(x + a) &= \log(a) + \frac{x}{a} - \frac{x^2}{2a^2} + \frac{x^3}{3a^3} \dots \\ &= \log(a) + \log\left(1 + \frac{x}{a}\right). \end{aligned} \quad (4.10)$$

i.e., taking the logarithm of speech plus additive noise as in an ASR front-end, then removing the constant via CMN yields something based on SNR.

I tried a quick experiment using SNR instead of spectral power and the results were extremely encouraging.

Further, something else was clear to me: Substituting  $\xi = \sigma/\nu$  into equation 4.9, it becomes:

$$p(t | \xi) = \int_0^{\infty} d\nu \frac{1}{\pi\nu(1 + \xi)} \exp\left(-\frac{|t|^2}{\nu(1 + \xi)}\right) \frac{B^A}{\Gamma(A)} \nu^{-A-1} \exp\left(-\frac{B}{\nu}\right) \quad (4.11)$$

$$= \frac{B^A}{\Gamma(A)} \frac{1}{\pi(1 + \xi)} \int_0^{\infty} d\nu \nu^{-A-2} \exp\left(-\frac{|t|^2 + B(1 + \xi)}{\nu(1 + \xi)}\right) \quad (4.12)$$

i.e., all the terms in  $\nu$  collect together. The noise marginalisation gives a tractable form!

This became the central concept of the work in Garner (2009a). In fact, that paper really has two contributions:

1. The first idea is that, in ASR, the variable that the engine is interested in is the variance,  $\sigma$ , rather than the undistorted observation,  $s$ , that is sought after in speech enhancement. Given that it is  $\sigma$  that is required usually, it is  $\sigma/v$  in the SNR case.
2. Given  $\xi = \sigma/v$ , it is possible to do a proper Bayesian analysis to infer the right variable. The analysis follows fairly trivially from the insight.

Aside from the marginalisation, the other aspect of the Bayesian approach is the prior. In Garner (2009a), I used a gamma distribution. In fact, this distribution came from earlier experiments reported in Garner (2009b). That it yielded a solution in the form of a cubic was not a problem because the same thing had occurred in Garner (2009b), and I had the code to solve it. The main reason to choose a gamma distribution was that the parameters ( $\beta$  at least) can be set using some knowledge of the overall SNR. At the time, I intended to try different priors; it is still a possible research direction.

#### 4.5.9 Features and databases

In turning the conference paper (Garner, 2009a) into a journal paper (Garner, 2011), the thing that most worried me was the database. I had done many experiments using Aurora-2, which is convenient, but has two major problems:

1. It is only digits; the grammar is simple and the phonetic coverage is small.
2. It is artificial data; noise is added in the computer.

Further, Idiap had funding to work on meeting data. This led to two unsuccessful and time consuming experiments:

1. I tried using the SNR features to train a meeting recogniser and test on rich text (the NIST RT07 data<sup>3</sup>). The result was worse than with normal features. In fact, this was predicted by the Aurora-2 results: meetings and rich text are not so noisy, and are matched conditions; SNR features do not perform well on the Aurora-2 equivalent.
2. I trained a recogniser on LDC (Linguistic Data Consortium) Wall Street Journal data (si-84) and tested on RT07. This was a mismatched system, but too mismatched. Error rates were in the 80s; far too large to infer anything useful.

Somewhat reluctantly, I turned to the other Aurora databases. Aurora-3 is still digits, but with real noise; Aurora-4 is continuous speech, albeit with artificially added noise and only 5000 word vocabulary. Here, however, results were encouraging.

In addition, I looked at PLP features. Although I had never used them before, they were popular at Idiap. SNR turned out to benefit PLPs more than MFCCs.

This was enough to write the journal paper. As such, the contribution of the journal version is mainly experimental; it demonstrates under what conditions SNR features work. Whilst it briefly discusses what does not work (the meeting room data), it is written from a positive point of view. In the context of this “critical analysis”, however, it is appropriate to talk about the negative results, hence the discussion above.

However, the journal version was also an opportunity to discuss other things. The main item was the relationship with the articulation index (AI).

---

<sup>3</sup>NIST: National Institute of Standards and Technology. RT: Rich transcription

### 4.5.10 Summary

The contributions over the three papers are:

**A decision theoretic framework for VAD** in which the otherwise arbitrary threshold is indicated by priors and costs.

**A differential VAD** where the likelihood from the spectrum is that of a high-pass filter.

**An SNR based feature** that is indicated by the use of cepstral normalisation.

## 4.6 With hindsight

The work in these three papers was done over a number of years. It represents my gradually learning how to better deal with the same underlying model. It follows that I would not do it the same way if I began again.

### 4.6.1 VAD

Of course, the VAD would benefit from the marginalisation discussed later. In fact, the VAD gain that prompted the work in the first place was probably due to the exponential form of the likelihood ratio. This in turn is because it is not a marginal estimate; the marginalisation tends to replace exponential forms with polynomial forms. In this sense, marginalising out the noise in the VAD likelihood ratio is likely to have a significant effect.

The noise tracker could also be improved significantly: There is a circular dependency in the VAD where the VAD depends on the noise estimate, which in turn depends on VAD. The minimum tracking that was used later would remove the dependency.

### 4.6.2 SNR features

One conclusion from Garner (2009a) is that introduction of a prior does not help so much. In fact, the form of the prior was probably wrong. If the prior were parameterised as  $1 + \xi$  instead of  $\xi$ , it is likely that it would be almost conjugate.

However, this is ongoing research, and these options all remain open.

### 4.6.3 Capacity of a Gaussian channel

My SNR feature and the articulation index (AI) have the same form as the capacity of a Gaussian channel, a standard result in information theory. This merits a short analysis:

The formulation below is essentially a worked solution of exercise 11.2 of MacKay (2003), but note that the notation for variances is almost the opposite of MacKay's. The capacity,  $C$ , of a channel is defined to be the maximum mutual information between the input,  $x$ , and output,  $y$  (being instantiations of the random variables  $X$  and  $Y$  respectively):

$$C = I(X; Y) = \int \int dx dy p(x, y) \log \frac{p(x, y)}{p(x) p(y)} \quad (4.13)$$

$$= \int \int dx dy p(y | x) p(x) \log \frac{p(y | x)}{p(y)} \quad (4.14)$$

$$= \int dx p(x) \int dy p(y | x) \log p(y | x) - \int dy p(y) \log p(y) \quad (4.15)$$



In the case of a Gaussian channel, it can be shown that the information throughput is maximised when the source is also Gaussian. In this case,

$$p(y|x) = \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(y-x)^2}{2\nu}\right) \quad (4.16)$$

and

$$p(y) = \frac{1}{\sqrt{2\pi(\sigma+\nu)}} \exp\left(-\frac{y^2}{2(\sigma+\nu)}\right) \quad (4.17)$$

The entropy terms are standard results and the mutual information is then

$$C = - \int dx p(x) \log \sqrt{2\pi e\nu} + \log \sqrt{2\pi e(\sigma+\nu)} \quad (4.18)$$

$$= \frac{1}{2} \log(\sigma+\nu) - \frac{1}{2} \log(\nu) \quad (4.19)$$

$$= \frac{1}{2} \log\left(1 + \frac{\sigma}{\nu}\right) \quad (4.20)$$

Notice that the final integral above is rendered trivial by the fact that entropy does not depend on the mean of a (Gaussian) distribution.

It follows that the SNR feature and the AI are actually measuring mutual information. Two things are evident:

1. Although the noisy channel is evident, there is no coding going on. In this sense, the capacity of the channel is not compromised beyond the additive noise, and the mutual information is the maximum possible — the channel capacity.
2. Notice that the above derivation relies on entropy being independent of the mean. The SNR feature also arises from the same operation — mean subtraction — that can be seen as converting it to mutual information.

So, the channel capacity and AI appear to be consistent.

## 4.7 Impact

VAD is an old subject; all labs have a different way of approaching it. The paper on VAD has just 3 citations according to Google scholar, perhaps owing to this kind of saturation. It was, however, patented by Canon Inc. (Garner et al., 2004a), and may even be used in their products.

As the later work is recent, it is difficult to measure impact; Google Scholar does not report any citations. Certainly the work is ongoing, however, and is the basis of at least one grant application.

It is perhaps worth noting that the final paper was well received at review, one reviewer going as far as writing:

The author has presented a novel way of looking at some standard front-end processing that I believe is quite insightful. It is rare to read a paper that casts a new perspective on something that is supposed to be understood. I am happy to recommend the paper for publication as I am sure many readers will find it thought provoking.

## Bibliography

- Yariv Ephraim and David Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-32(6):1109–1121, December 1984.
- ETSI. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. ETSI Standard 202 050, ETSI, 2002. V1.1.1.
- Philip Garner, Toshiaki Fukada, and Yasuhiro Komori. Signal detection using maximum a posteriori likelihood and noise spectral difference. United States Patent 7475012, Canon Kabushiki Kaisha, December 2004a. URL <http://www.google.com/patents?id=z9CWAAAAEBAJ>. Issued: January 6, 2009.
- Philip N. Garner. SNR features for automatic speech recognition. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, December 2009a.
- Philip N. Garner. A MAP approach to noise compensation of speech. Idiap-RR 08-2009, Idiap, June 2009b. URL [http://publications.idiap.ch/downloads/reports/2009/Garner\\_Idiap-RR-08-2009.pdf](http://publications.idiap.ch/downloads/reports/2009/Garner_Idiap-RR-08-2009.pdf).
- Philip N. Garner. Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition. *Speech Communication*, 53(8):991–1001, October 2011.
- Philip N. Garner, Toshiaki Fukada, and Yasuhiro Komori. A differential spectral voice activity detector. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, May 2004b.
- Guillaume Lathoud, Mathew Magimai-Doss, Bertrand Mesot, and Hervé Bouchard. Unsupervised spectral subtraction for noise-robust ASR. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, December 2005.
- David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- Pedro J. Moreno, Bhiksha Raj, and Richard M. Stern. A vector Taylor series approach for environment-independent speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 733–736, Atlanta, US, May 1996.
- Climent Nadeu, Dušan Macho, and Javier Hernando. Time & frequency filtering of filter bank energies for robust HMM speech recognition. *Speech Communication*, 34:93–114, April 2001.
- Jongseo Sohn and Wonyong Sung. A voice activity detector employing soft decision based noise spectrum adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 365–368, May 1998.
- Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, January 1999.

# Chapter 5

## Conclusions

### 5.1 Hypotheses

The papers in the previous chapters each contained their own conclusions. The chapters themselves also each contain concluding material. It remains to conclude the thesis as a whole in the sense of its binding topic of Bayesian approaches to uncertainty.

In the introduction, two hypotheses were stated:

1. Where an existing technique is somehow ad-hoc or not rigorous, we hypothesise that making it rigorous will lead to benefit in terms of allowing extensions that would not be possible otherwise.
2. Where an existing technique is rigorous, but not Bayesian, we hypothesise that making it Bayesian will lead to benefit in terms of robustness to small sample sizes.

Many of the papers addressed the first hypothesis. In the case of dialogue, it was shown that a rigorous formulation could generalise usefulness, leading to the multi-class scenario that was not possible with the basic two-class formulation. Further, the multi-class and model based formulation led to a rigorous metric to choose keywords that significantly outperformed other formulations.

In the case of formant analysis, the rigorous formulation not only allowed a coherent solution, but also led directly to being able to train the system. The training would not have been possible otherwise.

Finally, in the case of VAD, the statistical rigour existed already. However, casting a new idea (the high-pass filter) in the same rigorous manner allowed it to be integrated properly where it may not have been otherwise.

So, in at least three cases, the first hypothesis is demonstrated. Certainly there is no counter-example; no rigorous formulation led to inferior performance.

With regard to the second hypothesis, again there is evidence in its favour. Perhaps most persuasively, the dialogue scenario presented a requirement for inference based on few or even zero observations (words observed in one dialogue move might be absent in another). In this case the Bayesian formulation with explicit use of a prior led to a rigorous solution. In turn, the keyword selection strategy was also robust to small sample sizes, tending to reject singletons.

In the formant case, the Bayesian solution also led to a rigorous and stable result when formants were perhaps not even present. In this situation, the prior information was clear and required a subjective Bayesian view.

In the case of SNR features, a prior was shown to be beneficial, although perhaps not to the same extent as the other two cases.

So, the second hypothesis is also demonstrated. Again, there is certainly no counter-example.

## 5.2 Corollaries

Aside from the hypotheses, the results also lead to some corollaries. The first is that the pursuit of a rigorous solution can influence the underlying choice of model. This happened in each case: The word model was changed from multinomial to Poisson, the interpretation of the formant confidence was changed to be a variance, and the SNR feature was found to better suit the assumed model.

Although this change of model is quite tangible, there is also an intangible side to it. The pursuit of a good model leads to an understanding of the underlying mechanisms of the technology. In the SNR case in particular, it led to a relationship with the articulation index. Whilst I am hesitant to read too much into this, these observations are nevertheless interesting and merit further thought and even work.

Another corollary is that the form of the rigour is less important than the actual rigour. For instance, in the dialogue case, the fact that the formulation was Bayesian was important, but the form of the prior was not. The simple approximate prior was as good as the better fitting but more involved prior. The same was true of the SNR features: changing from uninformative to gamma did not improve results a great deal.

In fact, none of the results in this thesis represent game-changing improvements. Other authors have published better results on similar data suggesting that more involved models are necessary. Nevertheless, none of this is in conflict. The same rigorous method applied to more involved models ought to lead to similar benefits.

# Appendix A

## Full list of publications

Philip N. Garner. *Bayesian Approaches to Uncertainty in Speech Processing*. Phd by publication, School of Computing Sciences, University of East Anglia, March 2012.

Mohammad J. Taghizadeh, Philip N. Garner, and Hervé Bouchard. Microphones array beam-pattern characterization for hands-free speech applications. In *Proceedings of the Seventh IEEE Sensor Array and Multichannel Signal Processing Workshop*, Hoboken, NJ, USA, June 2012. To appear.

David Imseng, Hervé Bouchard, and Philip N. Garner. Boosting under-resourced speech recognizers by exploiting out of language data - case study on Afrikaans. In *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages*, Cape Town, South Africa, May 2012a. To appear.

David Imseng, Hervé Bouchard, and Philip N. Garner. Using KL-divergence and multilingual information to improve ASR for under-resourced languages. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012b. To appear.

Lakshmi Saheer, Junichi Yamagishi, Philip N. Garner, and John Dines. Combining vocal tract length normalization with hierarchical linear transformations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012. To appear.

Thomas Hain and Philip N. Garner. Speech recognition. In Steve Renals, Hervé Bouchard, Jean Carletta, and Andrei Popescu-Belis, editors, *Multimodal Signal Processing: Human Interactions in Meetings*, chapter 5. Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK, 2011.

Thomas Hain, Lukáš Burget, John Dines, Philip N. Garner, František Grézl, Asmaa El Hanani, Marijn Huijbregts, Martin Karafiát, Mike Lincoln, and Vincent Wan. Transcribing meetings with the AMIDA systems. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):486–498, February 2012.

Hervé Bouchard, John Dines, Mathew Magimai-Doss, Philip N. Garner, David Imseng, Petr Motlicek, Hui Liang, Lakshmi Saheer, and Fabio Valente. Current trends in multilingual

- speech processing. *Sādhanā*, 36(5):885–915, October 2011. Invited paper for special issue on the topic of Speech Communication and Signal Processing.
- Philip N. Garner. Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition. *Speech Communication*, 53(8):991–1001, October 2011.
- David Imseng, Hervé Boudouard, John Dines, Philip N. Garner, and Mathew Magimai.-Doss. Improving non-native ASR through stochastic multilingual phoneme space transformations. In *Proceedings of Interspeech*, Florence, Italy, August 2011.
- Andrei Popescu-Belis, Majid Yazdani, Alexandre Nanchen, and Philip N. Garner. A just-in-time retrieval system for dialogues or monologues. In *Proceedings of the 12th Annual SIG-Dial Meeting on Discourse and Dialogue*, pages 350–352, Portland, OR, USA, June 2011a.
- Andrei Popescu-Belis, Majid Yazdani, Alexandre Nanchen, and Philip N. Garner. A speech-based just-in-time retrieval system using semantic search. In *Proceedings of the ACL 2011 System Demonstrations*, pages 80–86, Portland, OR, USA, June 2011b.
- Mohammad J. Taghizadeh, Philip N. Garner, Hervé Boudouard, Hamid R. Abutalebi, and Asaei Afsaneh. An integrated framework for multi-channel multi-source localization and voice activity detection. In *Proceedings of The Third Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, Edinburgh, UK, May 2011.
- Mirjam Wester, John Dines, Matthew Gibson, Hui Liang, Yi-Jian Wu, Lakshmi Saheer, Simon King, Keiichiro Oura, Philip N. Garner, William Byrne, Yong Guan, Teemu Hirsimäki, Reima Karhila, Mikko Kurimo, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, and Junichi Yamagishi. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *Proceedings of the 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, September 2010.
- Lakshmi Saheer, John Dines, Philip N. Garner, and Hui Liang. Implementation of VTLN for statistical speech synthesis. In *Proceedings of the 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan, September 2010a.
- Philip N. Garner and John Dines. Tracter: A lightweight dataflow framework. In *Proceedings of Interspeech*, Makuhari, Japan, September 2010.
- Thomas Hain, Lukas Burget, John Dines, Philip N. Garner, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan. The AMIDA 2009 meeting transcription system. In *Proceedings of Interspeech*, Makuhari, Japan, September 2010.
- Danil Korchagin, Philip N. Garner, and Petr Motlicek. Hands free audio analysis from home entertainment. In *Proceedings of Interspeech*, Makuhari, Japan, September 2010a.
- Petr Motlicek, Fabio Valente, and Philip N. Garner. English spoken term detection in multilingual recordings. In *Proceedings of Interspeech*, Makuhari, Japan, September 2010.
- Afsaneh Asaei, Philip N. Garner, and Hervé Boudouard. Sparse component analysis for speech recognition in multi-speaker environment. In *Proceedings of Interspeech*, Makuhari, Japan, September 2010.

- Mikko Kurimo, William Byrne, John Dines, Philip N. Garner, Matthew Gibson, Yong Guan, Teemu Hirsimäki, Reima Karhila, Simon King, Hui Liang, Keiichiro Oura, Lakshmi Saheer, Matt Shannon, Sayaka Shiota, Jilei Tian, Keiichi Tokuda, Mirjam Wester, Yi-Jian Wu, and Junichi Yamagishi. Personalising speech-to-speech translation in the EMIME project. In *Proceedings of the ACL 2010 System Demonstrations*, pages 48–53, Uppsala, Sweden, July 2010.
- Danil Korchagin, Philip N. Garner, and John Dines. Automatic temporal alignment of AV data with confidence estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, USA, March 2010b.
- Lakshmi Saheer, Philip N. Garner, John Dines, and Hui Liang. VTLN adaptation for statistical speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, USA, March 2010b.
- Philip N. Garner. SNR features for automatic speech recognition. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, Merano, Italy, December 2009.
- Philip N. Garner, John Dines, Thomas Hain, Asmaa El Hannani, Martin Karafiát, Danil Korchagin, Mike Lincoln, Vincent Wan, and Le Zhang. Real-time ASR from meetings. In *Proceedings of Interspeech*, Brighton, UK, September 2009.
- Kenichi Kumatani, John McDonough, Barbara Rauch, Dietrich Klakow, Philip N. Garner, and Weifeng Li. Beamforming with a maximum negentropy criterion. *IEEE Transactions on Audio, Speech and Language Processing*, 17(5):994–1008, July 2009.
- Philip N. Garner. Silence models in weighted finite-state transducers. In *Proceedings of Interspeech*, Brisbane, Australia., September 2008.
- Kenichi Kumatani, John McDonough, Barbara Rauch, Philip Garner, John Dines, and Weifeng Li. Maximum kurtosis beamforming with the generalized sidelobe canceller. In *Proceedings of Interspeech*, Brisbane, Australia., September 2008a.
- Kenichi Kumatani, John McDonough, Dietrich Klakow, Philip N. Garner, and Weifeng Li. Adaptive beamforming with a maximum negentropy criterion. In *Proceedings of the Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Italy, May 2008b.
- Kenichi Kumatani, John McDonough, Stefan Schacht, Dietrich Klakow, Philip Garner, and Weifeng Li. Filter bank design based on minimization of individual aliasing terms for minimum mutual information subband adaptive beamforming. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, April 2008c.
- Philip N. Garner, Toshiaki Fukada, and Yasuhiro Komori. A differential spectral voice activity detector. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, May 2004.
- Jason P. A. Charlesworth and Philip N. Garner. Spoken content. In B. S. Manjunath, Philippe Salembier, and Thomas Sikora, editors, *Introduction to MPEG-7: Multimedia Content Description Interface*, chapter 18, pages 299–316. John Wiley & Sons Ltd., July

2002. URL <http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471486787,descCd-tableOfContents.html>.
- Philip N. Garner and Adam T. Lindsay, editors. *Information Technology - Multimedia Content Description Interface - Part 4: Audio*. Number 15938-4:2002. ISO/IEC, 2002. International Standard.
- Jason P. A. Charlesworth and Philip N. Garner. SpokenContent representation in MPEG-7. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6), June 2001. Special Issue on MPEG-7.
- J. P. A Charlesworth and P. N. Garner. Spoken content metadata and MPEG-7. In *Proceedings ACM Multimedia 2000 Workshops*, pages 81–84, Marina Del Rey, California, November 2000. ACM, PO Box 11405, New York, NY 10286 1405.
- Adam T. Lindsay, Savitha Srinivasan, Jason P. A. Charlesworth, Philip N. Garner, and Werner Kriechbaum. Representation and linking mechanisms for audio in MPEG-7. *Signal Processing: Image Communication*, 16:193–209, 2000.
- Andrew R. Webb and Philip N. Garner. A basis function approach to position estimation using microwave arrays. *Applied Statistics*, 48 part 2:197–209, 1999.
- Philip N. Garner and Wendy J. Holmes. On the robust incorporation of formant features into hidden Markov models for automatic speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, 1998.
- John N. Holmes, Wendy J. Holmes, and Philip N. Garner. Using formant frequencies in speech recognition. In *Proceedings of EUROSPEECH*, volume 4, pages 2083–2086, September 1997.
- Philip N. Garner. On topic identification and dialogue move recognition. *Computer Speech and Language*, 11:275–306, 1997.
- Philip N. Garner and Aidan Hemsworth. A keyword selection strategy for dialogue move recognition and multi-class topic identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 1997.
- Philip N. Garner, Sue R. Browning, Roger K. Moore, and Martin J. Russell. A theory of word frequencies and its application to dialogue move recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1880–1883, October 1996.
- Andrew R. Webb and Philip N. Garner. Source position estimation using radial basis functions. In *Proceedings 13th International Conference on Pattern Recognition*, volume IV, pages 3–7, Vienna, 1996.
- B. Steer, J. Kloske, P. Garner, L. LeBlanc, and S. Schock. Towards sonar based perception and modelling for unmanned untethered underwater vehicles. In *Proceedings IEEE International Conference on Robotics and Automation*, volume 2, pages 112–116, May 1993.



## Appendix B

# Letters from co-authors

The following pages comprise four letters from co-authors:

1. Roger Moore  
<mailto:r.k.moore@dcs.shef.ac.uk>
2. Aidan Hemsworth  
<mailto:ahemsworth@hotmail.com>
3. Wendy Holmes  
<mailto:w.holmes@aurix.com>
4. Toshiaki Fukada  
<mailto:fukada.toshiaki@canon.co.jp>

Department of Computer Science  
University of Sheffield  
Regent Court  
211 Portobello  
Sheffield S1 4DP  
UK

10<sup>th</sup> June 2011

**Re: Garner et al, 1996**

Dear Sir/Madam,

I am writing with reference to the following scientific paper ...

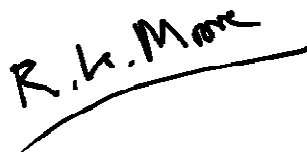
Garner, P. N., Browning, S. R., Moore, R. K., & Russell, M. J. (1996). A theory of word frequencies and its application to dialogue move recognition, *Int. Conf. on Spoken Language Processing* (pp. 1880-1883). Philadelphia.

As co-author of the above paper, and as Head of the UK Government's Speech Research Unit (SRU) at the time the relevant research was undertaken, I can confirm that the investigation was the prime responsibility of Mr. Phil Garner – a key member of the SRU at the time.

This particular research was the result of bringing together Mr. Garner's specialist skills in mathematical and statistical modelling with the challenges posed by contemporary spoken language processing. Mr. Garner was the main driving force behind the research, and he was personally responsible for conducting the investigation, analysing the outcomes, and publishing the results.

As I recall, the paper attracted a number of accolades at the time, and all credit was given to Mr. Garner as the prime author of the work.

Yours sincerely



**Prof. Roger K. Moore** BA(Hons) MSc PhD FIOA FISCA MIET MIEEE  
Personal Tutor

June 13, 2011

To whom it may concern

Dear Sir / Madam,

Philip N. Garner and Aidan Hemsworth. A keyword selection strategy for dialogue move recognition and multi-class topic identification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, April 1997.

The above paper was written whilst Phil and I were working at DERA in Malvern.

At the time, Phil had asked me to work on language modelling approaches to text classification. It was the same thing he was doing, but using different techniques and different databases. He defined the problem and suggested the solutions, but I did the implementation. Reciprocally, he had already implemented the Poisson distribution based approach.

I started to investigate information theoretic approaches to word selection; Phil also came up with and implemented a Bayesian means to do the same thing.

The paper then arose naturally as a comparison of the different approaches. I ran my information theoretic and language modelling techniques on the dialogue database he was using; he ran his Poisson based techniques on the LOB database I was using. The analysis of the outcome was trivial because we were using the same comparison metrics.

The paper was written by Phil because he had the better understanding of the two techniques, and had written such papers before.

Sincerely,

A handwritten signature in black ink, appearing to read 'Aidan Hemsworth', written in a cursive style.

Aidan Hemsworth



**Aurix Limited**  
Malvern Hills Science Park  
Geraldine Road  
Malvern  
Worcestershire  
WR14 3SZ  
United Kingdom

w.holmes@aurix.com

Tel 01684 585 119

Fax 01684 585 151

16<sup>th</sup> August 2011

To whom it may concern

Dear Sir/Madam,

I am writing with reference to the following scientific papers:

J. N. Holmes, W. J. Holmes and P. N. Garner (1997) "Using formant frequencies in speech recognition", *Proc. EUROSPEECH'97*, Rhodes, pp. 2083-2086.

P. N. Garner and W. J. Holmes (1998) "On the robust incorporation of formant features into hidden Markov models for automatic speech recognition", *Proc. IEEE ICASSP*, Seattle, pp. 1-4.

At the time that these papers were written, both Phil Garner and I were working at the UK Government's Speech Research Unit.

The research focussed on novel methods for using formant features in HMM-based automatic speech recognition, with the emphasis being on developing techniques that were robust to the types of errors normally associated with extracting and using formant information. The formant analyser was the invention of my father, John Holmes, and I carried out the experimental work. Phil developed the novel mathematical techniques on which the work was based.

Phil's major contributions to the research were in identifying that the notion of 'confidence' in formant measurement could be expressed mathematically as a variance, and then in deriving the associated mathematics for applying this concept in both recognition and training of HMMs.

Phil contributed to the writing of the EUROSPEECH paper, especially with the mathematical foundations for the techniques. He both wrote and presented the more mathematical ICASSP paper.

Yours sincerely,

Wendy Holmes BSc MPhil PhD

Principal Consultant

June 15, 2011

Applied Software Technology Development Center  
Canon Inc.  
30-2, Shimomaruko 3-chome, Ohta-ku,  
Tokyo 146-8501  
Japan

To whom it may concern

Dear Sir / Madam,

Philip N. Garner, Toshiaki Fukada, and Yasuhiro Komori. A differential spectral voice activity detector. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Montreal, May 2004.

The above paper was written while Phil was here at Canon Inc. in Tokyo. I was guiding the research, along with Dr. Komori the group manager. With regard to the points in the guideline:

The investigation was designed jointly, but followed on from some previous work in voice activity detection done in the group. For instance, the database had been collected already, and we had initial evaluations. The intention was to improve on the VAD that we were using at the time; Phil had been investigating alternatives. The differential idea was mine, but one of many ideas that came up during discussions.

All of the mathematical derivation and programming work was done by Phil, as were the experiments and analysis of results. The "Gross error" measure had been used previously; Phil formalized it into the table in the paper.

As for the publication, Phil wrote the paper then Dr. Komori and I acted essentially as reviewers. This was natural in the sense that Phil is native English. Invention based on this research was also applied to the US and Japan Patent Office and granted as patents (USP7475012, JP04497911), respectively.

Yours sincerely,



Toshiaki Fukada, Ph.D.

Manager

# Appendix C

## Papers

The remaining pages comprise the eight papers summarised in the document.

# A THEORY OF WORD FREQUENCIES AND ITS APPLICATION TO DIALOGUE MOVE RECOGNITION

*P. N. Garner, S. R. Browning, R. K. Moore and M. J. Russell*

Defence Research Agency, St Andrews Rd, Malvern, WORCS. WR14 3PS, UK

Email: [garner@signal.dra.hmg.gb](mailto:garner@signal.dra.hmg.gb)

©British Crown Copyright 1996/DERA

Published with the permission of the Controller of Her Britannic Majesty's Stationery Office.

## ABSTRACT

Dialogue move recognition is taken as being representative of a class of spoken language applications where inference about high level semantic meaning is required from lower level acoustic, phonetic or word based features. Topic identification is another such application. In the particular case of inference from words, the multinomial distribution is shown to be inadequate for modelling word frequencies, and the multivariate Poisson is a more reasonable choice. Zipf's law is used to model a prior distribution. This more rigorous mathematical formulation is shown to improve dialogue move classification both subjectively and quantitatively.

## 1. INTRODUCTION

It has been suggested [5] that a dialogue, that is, the interaction between two or more people in a conversation, can be represented as a series of moves (as a game of chess consists of alternate moves). These moves follow a natural sequence, with alternatives and counter moves. The dialogue moves dictate portions of speech that can be classified into the different move types, and may in turn dictate sensible bounds between which processing can be carried out.

The dialogue moves also form a natural part of the progression from raw acoustic data to natural language processing. Inference can proceed in either direction: down towards the acoustic recogniser or up towards the natural language processor. This paper is concerned with the latter, and in particular with the question of whether it may be possible to construct a data driven natural language processor. Dialogue move recognition can be viewed as a metric against which the contribution of dialogue moves to natural language processing can be judged.

## 2. AN INITIAL EXPERIMENT

### 2.1. Data

The HCRC map task corpus [1] has been annotated at the dialogue move level, and this database was used as an experimental vehicle. Only utterances which could be identified as

belonging to one move category were used, and all non-word annotation was stripped out. Punctuation was removed, and upper case letters were converted to lower case.

The 128 dialogues were then split into training and testing sets of 64 dialogues each such that no map appeared in both sets. This was to prevent discrimination occurring on particular map features, hence forcing the use of other words more indicative of semantic meaning. The training and testing sets contained 11799 utterances and 10265 utterances respectively.

### 2.2. Methodology

The methodology was essentially that used in word based topic identification, outlined as follows:

The moves were assumed to be samples from a random variable  $\mathcal{M} \in \{m_1, m_2, \dots, m_M\}$ ; in this case, the number of possible moves,  $M$ , was 12. Given an utterance  $x$ , and training data  $D$ , the problem is to maximise the likelihood of the move  $m_i$ . Using Bayes's theorem,

$$P(\mathcal{M} = m_i | x, D) = \frac{P(x | \mathcal{M} = m_i, D) P(\mathcal{M} = m_i | D)}{P(x | D)}$$

The denominator,  $P(x | D)$ , is independent of the move and can be ignored.

Assuming  $P(m_i)$  to be an abbreviation for  $P(\mathcal{M} = m_i)$ ,  $P(m_i | D)$  is the prior (prior to the utterance but posterior to the data), and was calculated as the number of moves of type  $m_i$  in  $D$  divided by the total number of moves in  $D$ .

$P(x | m_i, D)$  is the likelihood. Here, it was assumed that  $x$  was generated by sequentially sampling from a random variable  $\mathcal{W} \in \{w_1, w_2, \dots, w_V\}$ , where  $V$  is the vocabulary of the task, and samples from  $\mathcal{W}$  are independent. Hence, if  $x$  is  $K$  words in length,

$$\begin{aligned} P(x | m_i, D) &= P(\mathcal{W} = w_1, \mathcal{W} = w_2, \dots, \mathcal{W} = w_K | m_i, D), \\ &= P(w_1 | m_i, D) \\ &\quad \times P(w_2 | m_i, D) \times \dots \\ &\quad \times P(w_K | m_i, D). \end{aligned}$$

$P(w_k|m_i, D)$  was calculated as the number of words of type  $w_k$  in move  $m_i$  in  $D$  divided by the total number of words in move  $m_i$  in  $D$ . Where the count for a word in  $x$  was zero, that word was assumed to have occurred 0.5 times.

### 2.3. Results

Table 1 shows a confusion matrix for the classification problem so far described. The overall accuracy is 47.22%, and assuming the test set accuracy is binomially distributed [2], the 95% confidence limits for 10265 independent testing samples are around  $\pm 1\%$ .

Note that a disproportionate number of utterances have been classified as 'Ready'. This is counter intuitive; one would expect utterances about which the system was unsure to be classified as 'Acknowledge', since that is the most frequent class. Further, 'Acknowledge', 'Ready' and 'Reply-Y' are all basically affirmative utterances ("yes"), and one would expect them to be indistinguishable at this level.

## 3. PROBABILITY DISTRIBUTIONS

### 3.1. The Multinomial

When probabilities are calculated as a relative frequency as described, one is implicitly assuming a multinomial (dice throwing) distribution. That is, if the number of words of type  $w_i$  in a move is  $n_i$ , and  $N = \sum_{i=1}^V n_i$ , then  $P(w_i|D) = n_i/N$ . In fact, this is the maximum likelihood estimator of the true probability; it becomes more accurate as  $N \rightarrow \infty$ . In this case, though, some of the  $n_i$  are actually zero and the maximum likelihood estimator breaks down completely.

More light can be shed on the situation by considering a Bayesian formulation of the word probability problem [4]. Using a multinomial distribution with a flat Dirichlet prior, the probability of a single word  $w_i$  being drawn from  $\mathcal{W}$  is

$$P(w_i|D) = \frac{n_i + 1}{N + V}.$$

The formula now depends on  $V$ , the vocabulary of the task. This can be thought of intuitively too: Given a biased die, but no data upon which to base an approximation, most people would agree that a good starting point would be to assume a probability of throwing any particular number to be 1/6. This is implicitly based on the prior knowledge that a die has 6 sides.

This explains the reason for assuming  $n_i = 0.5$  for unseen words: the probability for  $n_i = 0$  is half that for  $n_i = 1$ .  $V$  is large, though, and whilst it is unknown it suggests that the maximum likelihood estimate is consistently an overestimate of the true posterior probability. The largest overestimates of this word probability will occur in the class for which  $N$  is smallest; the least frequent class is 'Ready'.

### 3.2. The Multivariate Poisson

If the underlying probability of drawing word  $w_i$  from  $\mathcal{W}$  is  $\omega_i$ , then the multinomial distribution is

$$P(\mathbf{n}|\boldsymbol{\omega}) = \frac{N!}{n_1! \dots n_V!} \omega_1^{n_1} \dots \omega_V^{n_V}$$

where  $\mathbf{n} = \{n_1, n_2, \dots, n_V\}$  and  $\boldsymbol{\omega} = \{\omega_1, \omega_2, \dots, \omega_V\}$ . Consider what would happen if this model were used to generate an infinite amount of data: It can be proved that if the  $\omega_i$  are constrained to be small enough such that  $N\omega_i \rightarrow \lambda_i$  as  $N \rightarrow \infty$ ,

$$P(\mathbf{n}|\boldsymbol{\lambda}) = \frac{\lambda_1^{n_1} \lambda_2^{n_2} \dots \lambda_{V-1}^{n_{V-1}}}{n_1! n_2! \dots n_{V-1}!} e^{-\lambda_1 - \lambda_2 - \dots - \lambda_{V-1}},$$

where  $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \dots, \lambda_{V-1}\}$ . This is the multivariate Poisson distribution.

Note that one of the  $\omega$  terms has disappeared. More correctly, any of the  $\omega$  terms can be made to disappear by simply grouping them into one term; the useful approach is to group all unknown words into a single  $\omega$ , and have that disappear. The result is a distribution which is independent of vocabulary; indeed it can be tailored to any arbitrarily sized vocabulary.

The intuitive approach to the above derivation is to consider several throws of a die.  $\omega_i$  relates to each individual throw, whereas  $\lambda_i$  is concerned with the rate of occurrence of the feature of interest.

The probability of an utterance of  $K$  words in length using a multivariate Poisson distribution and a gamma prior can be shown [4] to be

$$P(\mathbf{x}|D) = \prod_{i=1}^W \left( \frac{(N + \beta)^{n_i + \alpha}}{(N + \beta + K)^{n_i + \alpha + x_i}} \frac{\Gamma(n_i + \alpha + x_i)}{\Gamma(n_i + \alpha)} \right),$$

where  $n_i$  and  $N$  are the same as in the multinomial,  $x_i$  is the number of words of type  $w_i$  in  $x$ ,  $W$  is the number of 'keywords' and  $\alpha$  and  $\beta$  are the parameters of the gamma prior. Note that this calculation refers to the probability of the whole utterance, not the product of the probabilities of the individual words.

## 4. PRIOR INFORMATION

### 4.1. Zipf's Law

Whilst it is convenient to attach a flat prior to a distribution and simply let the data decide what to do, it must be acknowledged that prior information exists in the form of Zipf's law [7]. Zipf's law itself is an empirical law relating relative frequencies. If a graph is plotted of frequency as ordinate, and the words rank ordered on the abscissa, that is, the most frequent word on the left and the least frequent on the right, the points will form a smooth curve with approximately reciprocal square root form; the actual analytical



form is discussed by McNeil[6]. Further, this law will hold no matter which database is used.

Such a graph is not very useful in that form, but integrating up the vertical axis produces a graph which, suitably normalised, can be interpreted as 'Probability of Frequency', which in turn is the prior on the  $\lambda$  terms in the Poisson distribution. This is illustrated in figure 1, where the graph on the left is a traditional Zipf plot, and the one on the right is modified as described.

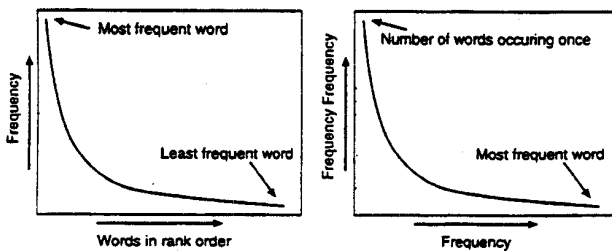


Figure 1: The Zipf plot, and how to modify it to relate to probability.

The graph on the right of figure 1 can be estimated with a histogram from a large dataset, and this is depicted in figure 2. The scatter plots refer to the King James version of the Bible, the entire radio 4 weather forecast spotting database [3], and the entire HCRC Map Task corpus. Two things are apparent from this plot:

1. All the plots are straight lines with the same gradient. If they are indeed the same, then Zipf's law holds, and one dataset can be used as a prior for another.
2. The fact that they are straight lines on a double logarithmic scale implies that the real curve is of the form  $y = Ax^m$ , where  $A$  is some normalising term and  $m$  is the gradient of the line.

Note that the map Task plot is only shown for reference. This is supposed to be prior information, and looking at any of the Map Task data is cheating, never mind looking at all of it.

The gamma distribution has a  $x^m$  term, so it ought to be possible to fit a gamma distribution to this database. The lines on Figure 2 illustrate this. The line labelled 'Gamma 1' is a gamma distribution with parameters  $\alpha = 0.1$  and  $\beta = 1$ ; 'Gamma 2' is the same with  $\beta = 10$ . Shrinking  $\alpha$  any more has the effect of moving the whole line downwards.

There is clearly nothing to be gained from setting  $\beta$  to be anything other than 0. It only acts as a prior on the number of observations, which is of the order of several thousand. Even a value of 10 introduces more curvature than can be justified. Setting  $\alpha$  to some small value may clearly be of benefit though.

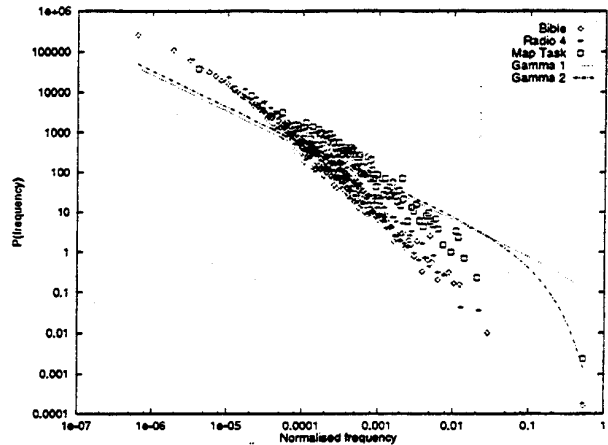


Figure 2: Modified Zipf plot for various data sources, with approximate gamma distribution fits.

## 5. EVALUATION

Table 2 shows a confusion matrix for the classification experiment using the Poisson based estimate with a gamma prior with  $\alpha$  set to 0.1. The classification rate is better than the maximum likelihood case; but more importantly, the misclassifications are much better distributed. No one class seems to mop up the ambiguous observations in a disproportionate manner. In fact, nothing is classified as 'Ready', but that is understandable since that category is indistinguishable from 'Acknowledge'.

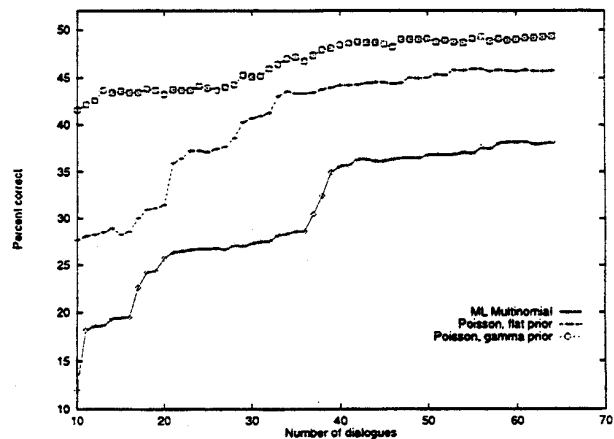


Figure 3: Classification rate as a function of amount of training data.

To evaluate the performance of the Poisson technique more fully, a test data set was constructed by randomly sampling 100 observations of each category from the test data previously described. With the classifier suitably modified for equal class membership priors, experiments were performed on training set sizes ranging from 10 to 64 dialogues. The results are shown in figure 3, confidence limits for 1200 test

samples are around  $\pm 3\%$ . This plot is very gratifying, showing that the Poisson based estimate performs better than the maximum likelihood multinomial, and that incorporation of a Zipf's law based prior further improves performance, especially for small amounts of training data.

## 6. CONCLUSIONS

It has been shown that the multivariate Poisson distribution is a justifiable and more suitable distribution to model word frequencies for dialogue move recognition. Incorporation of Zipf's law as a prior follows naturally and further improves performance.

Dialogue moves can be inferred from their constituent words to an accuracy of around 50% using a very simple unigram model, implying that better performance should be possible using a more involved N-gram Markov model.

corpus. *Language and Speech*, 34(4):351-366, 1991.

2. Mark D. Bedworth. On the quality and quantity of data and pattern recognition. Memorandum (unpublished), Defence Research Agency, St Andrews Rd, Malvern, WORCS, WR14 3PS, UK, 1992.
3. Mike J. Carey and E. S. Parris. Topic spotting using task independent models. In *Proceedings Eurospeech 95, Madrid*, pages 2133-2137, 1995.
4. Philip N. Garner. On topic spotting and dialogue move recognition. Memorandum (unpublished), Defence Research Agency, St Andrews Rd, Malvern, WORCS, WR14 3PS, UK, 1996.
5. Jaqueline C. Kowtko, Stephen D. Isard, and Gwyneth M. Doherty. Conversational games within dialogue. Technical report, Human Communication Research Centre, University of Edinburgh, 2 Buccleugh Place, Edinburgh EH8 9LW SCOTLAND, November 1993.

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1795	17	32	2	17	66	4	18	119	61	5	323	2459
Align	390	125	19	11	6	33	14	28	114	3	9	8	760
Check	29	38	273	38	46	251	40	40	209	21	37	15	1037
Clarify	7	11	54	35	15	135	7	5	111	8	31	4	423
Explain	23	23	52	15	172	43	11	9	277	82	77	2	786
Instruct	10	30	122	159	35	639	41	20	425	11	49	2	1543
Query-W	5	8	19	4	5	29	186	11	47	1	0	0	315
Query-YN	3	28	47	10	28	36	29	401	144	12	13	3	754
Ready	82	0	4	0	1	11	0	0	9	0	0	0	107
Reply-N	3	1	4	1	1	3	0	1	4	301	3	0	322
Reply-W	11	13	45	20	28	82	10	10	108	21	51	4	403
Reply-Y	329	14	25	4	21	32	3	14	35	11	8	860	1356
Total	2687	308	696	299	375	1360	345	557	1602	532	283	1221	10265

Table 1: Confusion matrix for the initial experiment, Accuracy = 47.22%

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1851	25	39	2	37	86	4	23	1	58	9	324	2459
Align	397	171	28	9	24	59	14	38	0	3	9	8	760
Check	41	42	326	28	109	359	28	53	0	11	23	17	1037
Clarify	12	13	69	28	37	212	4	9	0	4	30	5	423
Explain	42	37	101	12	379	86	9	23	0	35	58	4	786
Instruct	21	36	164	74	88	1052	27	34	0	6	39	2	1543
Query-W	9	15	34	3	9	39	187	17	0	0	2	0	315
Query-YN	12	32	70	3	74	81	25	438	0	3	13	3	754
Ready	87	1	4	0	2	12	0	0	0	0	1	0	107
Reply-N	6	1	8	1	10	3	0	1	0	289	3	0	322
Reply-W	22	18	56	16	83	130	6	14	0	9	44	5	403
Reply-Y	343	15	32	2	40	38	3	14	0	3	9	857	1356
Total	2843	406	931	178	892	2157	307	664	1	421	240	1225	10265

Table 2: Confusion matrix for the Poisson based classification, Accuracy = 54.77%

## 7. REFERENCES

1. Anne H. Anderson, Miles Bader, Ellen Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jaqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, and Henry S. Thompson. The HCRC map task
6. Donald R. McNeil. Estimating an author's vocabulary. *Journal of the American Statistical Association*, 68(341):92-96, March 1973.
7. G. K. Zipf. *The Psycho-Biology of Language*. Houghton-Mifflin, Boston, 1935.

# A KEYWORD SELECTION STRATEGY FOR DIALOGUE MOVE RECOGNITION AND MULTI-CLASS TOPIC IDENTIFICATION

Philip N. Garner and Aidan Hemsworth

Defence Research Agency, St Andrews Rd, Malvern, WORCS. WR14 3PS, UK  
Email: [garner@signal.dra.hmg.gb](mailto:garner@signal.dra.hmg.gb)

## ABSTRACT

The concept of usefulness for keyword selection in topic identification problems is reformulated and extended to the multi-class domain. The derivation is shown to be a generalisation of that for the two class problem. The technique is applied to both multinomial and Poisson based estimates of word probability, and shown to outperform or compare favourably to various information theoretic techniques classifying dialogue moves in the map task corpus, and reports in the LOB corpus.

## 1. INTRODUCTION

Over the past few years, a general class of problem has arisen where inference is required about high level semantic meaning from some lower level feature set. The main manifestation of this problem is in topic identification, where a system is required to detect when a 'Wanted' topic is being discussed in a stream of largely 'Unwanted' material. The source data can be text, the word level output of a speech recogniser [1], or acoustic or phonetic level data, for instance [2].

Topic identification is traditionally a two class problem, but can easily be extended to multi-class by partitioning the 'Wanted' class into sub-classes, for example [3]. The same methods have been used to do dialogue move recognition by other authors, eg. [4] and [5]; here the problem is specified in terms of spoken language understanding, but the methodology is exactly the same as in topic identification.

In all of these problems, one approach is to identify a set of 'keywords' or 'key features'  $\mathcal{W} = \{w_1, w_2, \dots, w_K\}$ , which are sufficient to distinguish the chosen classes. This reduced dictionary is then used to build language models each indicative of a particular class; the number of key features (dictionary size) is a trade off between complexity and performance. In the two class case, the decision rule is to assign the observation,  $\mathbf{x} = w_1, w_2, \dots, w_K$ , to the Wanted class,  $C_W$ , iff

$$\prod_{k=1}^K \frac{P(w_k|C_W)}{P(w_k|C_U)} > \lambda,$$

where the  $w_k$  are the independent constituent features of the observation, and  $\lambda$  is some threshold. In this paper, the features are words.

The metric dictating the choice of features follows directly from the decision rule: choose features which maximise the probability ratio inside the product (weighted by

the frequency of occurrence of those features). For this reason, this weighted ratio has been termed 'Usefulness' [6].

The decision rule in the multi-class case is more complex. If the set of  $M$  classes is  $\mathcal{M} = \{m_1, m_2, \dots, m_M\}$ , then the decision rule is to maximise

$$\max_i \frac{P(\mathbf{x}|m_i)P(m_i)}{P(\mathbf{x})}.$$

It is clear that a simple inequality cannot be formed resulting in a simple ratio.

## 2. INFORMATION THEORETIC MEASURES

It is reasonable to assume that keywords should be chosen which maximise some measure of information. Less clear, though, is which measure; three possible measures can be identified as follows.

Quoting Gallager [9], if  $m$  is a sample from  $\mathcal{M}$  and  $w$  is a sample from  $\mathcal{W}$ , the information provided about the event  $m = m_i$  by the occurrence of the event  $w = w_k$  is

$$I(m_i; w_k) = \log \frac{P(m_i|w_k)}{P(m_i)}.$$

This is the mutual information between the two events. To extend the measure to apply over all classes, consider the expectation over classes:

$$I(\mathcal{M}; w_k) = \sum_{i=1}^M \log \frac{P(m_i|w_k)}{P(m_i)} P(m_i).$$

Mutual information expressed in this way is very similar to the expression for the change in entropy (with one changed term):

$$I_E(\mathcal{M}; w_k) = - \sum_{i=1}^M P(m_i) \log P(m_i) + \sum_{i=1}^M P(m_i|w_k) \log P(m_i|w_k).$$

This has the intuitively appealing quality of representing the increase in entropy of the ensemble  $\mathcal{M}$  when word  $w_k$  is observed.

Salience has been used by Gorin [8] to rank words in order of importance to classify actions in a dialogue management system. Salience is defined as

$$S(\mathcal{M}; w_k) = \sum_{i=1}^M P(m_i|w_k)I(m_i; w_k).$$

Writing the three measures  $I(\mathcal{M}; w_k)$ ,  $I_E(\mathcal{M}; w_k)$  and  $S(\mathcal{M}; w_k)$ , which shall be referred to as mutual information, entropy and salience respectively, as

$$\begin{aligned} & \sum_{i=1}^M P(m_i) \log P(m_i|w_k) - \sum_{i=1}^M P(m_i) \log P(m_i) \\ & \sum_{i=1}^M P(m_i|w_k) \log P(m_i|w_k) - \sum_{i=1}^M P(m_i) \log P(m_i) \\ & \sum_{i=1}^M P(m_i|w_k) \log P(m_i|w_k) - \sum_{i=1}^M P(m_i|w_k) \log P(m_i), \end{aligned}$$

it is clear that they are intimately related, the only difference being whether the raw information term (the logarithm term) is weighted by  $P(m_i)$  or  $P(m_i|w_k)$ .

Gorin [8] uses some standard smoothed relative frequencies to estimate the probabilities above. In this paper, we use the maximum likelihood estimate

$$P(m_i) = \frac{n_i}{N},$$

where  $n_i$  is the number of occurrences of class  $m_i$  in the training data, and  $N$  is the total number of occurrences. The posterior measure  $P(m_i|w_k)$  is evaluated via Bayes's theorem:

$$P(m_i|w_k) = \frac{P(w_k|m_i)P(m_i)}{\sum_{i=1}^M P(w_k|m_i)P(m_i)}.$$

### 3. USEFULNESS

The decision rule itself can also indicate a measure of 'usefulness' for each possible word: The multi-class decision rule is to maximise

$$P(m_i|\mathbf{x}) = \frac{P(\mathbf{x}|m_i)P(m_i)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|m_i)P(m_i)}{\sum_{j=1}^M P(\mathbf{x}|m_j)P(m_j)}.$$

Denoting the reciprocal of this expression by  $\mathcal{P}_i$ , the problem is the same as minimising

$$\begin{aligned} \mathcal{P}_i &= \frac{P(\mathbf{x}|m_1)P(m_1)}{P(\mathbf{x}|m_i)P(m_i)} + \frac{P(\mathbf{x}|m_2)P(m_2)}{P(\mathbf{x}|m_i)P(m_i)} + \\ &\dots + \frac{P(\mathbf{x}|m_M)P(m_M)}{P(\mathbf{x}|m_i)P(m_i)}, \end{aligned}$$

which consists of easily differentiable parts. It is reasonable to assume that discriminative keywords will be those which lead to a high rate of change of this probability. Consider the expected rate of change of  $\mathcal{P}_i$  when a new feature or word is considered: By definition,

$$E\left(\frac{\partial \mathcal{P}_i}{\partial x_k}\right) = \sum_{k=1}^w \frac{\partial \mathcal{P}_i}{\partial x_k} P(w_k|m_i),$$

where there are  $x_k$  words of type  $w_k$  in  $\mathbf{x}$ . The new feature is unknown, and this is accounted for by integrating over all possible features. The features or words which have maximum effect upon the decision rule are those which minimise this expectation (largest negative value). It is clear that the most useful words are those which minimise

$$\frac{\partial \mathcal{P}_i}{\partial x_k} P(w_k|m_i).$$

This can be evaluated with all the  $x_k = 0$ , embodying the assumption that the usefulness of the occurrence of a word is independent of the number of times it has occurred already.

Thus far, the theory only addresses choosing keywords to discriminate one class from the others. A natural extension is to integrate over all classes:

$$E\left(\frac{\partial \mathcal{P}}{\partial x_k}\right) = \sum_{i=1}^M E\left(\frac{\partial \mathcal{P}_i}{\partial x_k}\right) P(m_i).$$

This is actually slightly non-intuitive in that a change in probability of one class will be accompanied by an opposite change in that of other classes. One might feel happier adding squared rates of change to capture both large positive and negative gradients, but in practice this makes little difference.

If it is assumed that the underlying model for the word generation is a multinomial (dice throwing) distribution, the probability of a sequence of words  $\mathbf{x}$  conditioned on the class, in a maximum likelihood sense, is

$$P(\mathbf{x}|m_i) = \prod_{k=1}^K \frac{n_{ik}}{D_i},$$

where there are  $n_{ik}$  words of type  $w_k$  and  $D_i$  words in total in class  $m_i$  of the training set. If  $U(w_k)$  is defined to be the usefulness of word  $w_k$ , then this results in a usefulness for word  $w_k$  of

$$U(w_k) = \sum_{i=1}^M \frac{n_{ik}}{D_i} \frac{n_i}{N} \sum_{\substack{j=1 \\ j \neq i}}^M \frac{n_j}{n_i} \log \frac{n_{jk}D_i}{n_{ik}D_j},$$

where there are  $n_j$  examples of class  $m_j$  in the training data. In practice, the two  $n_i$  terms cancel, and the  $N$  is unnecessary. In the special case of two classes, this expression can be written

$$\begin{aligned} U(w_k) &= -P(m_2)P(w_k|m_1) \log \frac{P(w_k|m_1)}{P(w_k|m_2)} \\ &\quad -P(m_1)P(w_k|m_2) \log \frac{P(w_k|m_2)}{P(w_k|m_1)}. \end{aligned}$$

Each of these terms is exactly the same as that given by [6], though from a much more general view, and corresponds to combining features indicative of the wanted class with features indicative of the unwanted class. For this reason, we feel justified in retaining the name usefulness. Curiously though, the term corresponding to class 1 is weighted by the probability of class 2 and vice-versa.

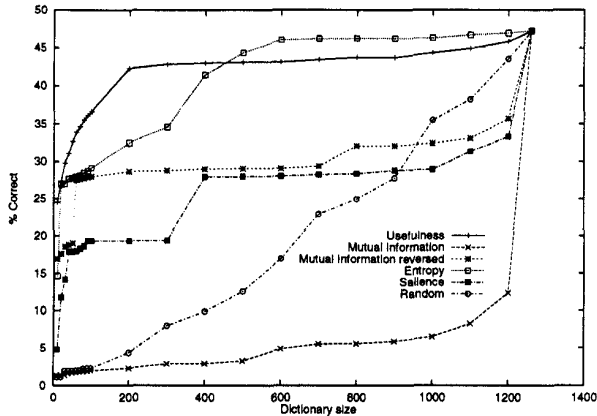


Figure 1: Map task corpus, multinomial

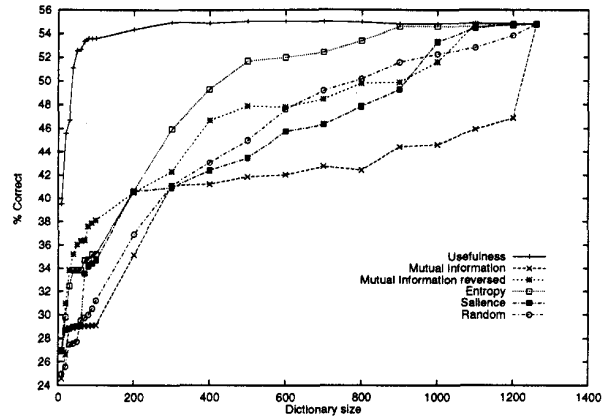


Figure 3: Map task corpus, poisson based

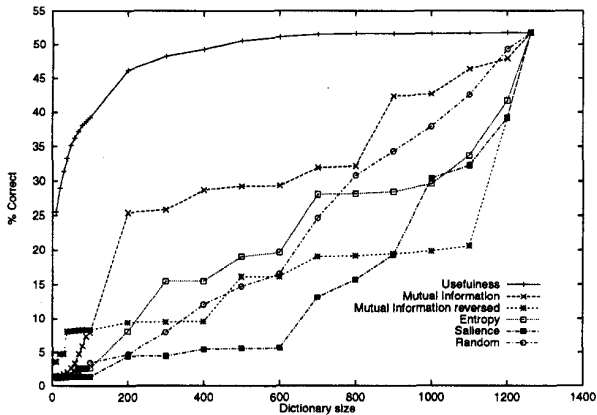


Figure 2: Map task corpus, absolute discounting

If it is assumed that the underlying model of word generation is Poisson, then from [5], the probability of the sentence is

$$P(\mathbf{x}|m_i) = \prod_{k=1}^W \left[ \frac{\Gamma(n_{ik} + \alpha + x_k)}{\Gamma(n_{ik} + \alpha)} \frac{(D_i + \beta)^{n_{ik} + \alpha}}{(D_i + \beta + K)^{n_{ik} + \alpha + x_k}} \right],$$

where there are  $W$  distinct words in the vocabulary, and  $\alpha$  and  $\beta$  are priors. By the same method as above, this results in a usefulness for word  $w_k$  of

$$U(w_k) = \sum_{i=1}^M P(x_k|m_i) \sum_{\substack{j=1 \\ j \neq i}}^M n_j [\log(D_i + \beta) - \log(D_j + \beta) + \psi(n_{jk} + \alpha) - \psi(n_{ik} + \alpha)],$$

where  $P(x_k|m_i)$  is the the probability of a sentence consisting of the single word  $w_k$ , and  $\psi$  is the digamma function.

#### 4. EXPERIMENTS

Two corpora were used: The HCRC Map Task Corpus [7], which is annotated at the dialogue move level, and the LOB

corpus, which is divided into reports and essays classified into different topics. Each corpus was stripped of punctuation and annotation, and translated entirely to lower case. The map task corpus was split into training and testing sets of 64 dialogues each such that no map occurred in both sets; this was to bias the discrimination against particular map features. There were 11799 moves in the training set and 10265 in the testing set. The LOB corpus was split by alternating reports into the training and testing sets; the training and testing sets both consisted of 250 reports.

Classification experiments were performed using language models built from both Poisson based and multinomial based probability measures, and classification rate was plotted against dictionary size for various keyword selection methods. Each probability measure was also tested against three randomly ordered dictionaries, the results of which were averaged to provide a baseline.

For the multinomial, out of vocabulary (OOV) words were handled in two different ways. The first, after Nowell [2], involved simply scoring OOV words as if they had occurred 0.5 times. The second was to use absolute discounting (for example [10]) to provide a smoothed estimate of word probabilities; this was only optimised for the largest dictionary size. In the Poisson based case, the hyperparameters  $\alpha$  and  $\beta$  were set to 0.1 and 0 respectively after [5]. The experimental results are shown in figures 1-5 (Note the different ordinate scales), except those for the basic multinomial on the LOB corpus, which scored consistently below 14%, and were omitted after space considerations.

#### 5. DISCUSSION

The Poisson based probability measure was developed specifically for this type of problem, indeed specifically to alleviate the OOV problems of the multinomial. It is gratifying, therefore, that the Poisson measure performs a good 5% better than the multinomial on the map task, and even better on the LOB corpus. In turn, the multi-class usefulness measure was developed specifically to complement the Poisson based probability, and performs consistently better than any other dictionary pruning method for the Poisson.

The comparative results are still informative though. In

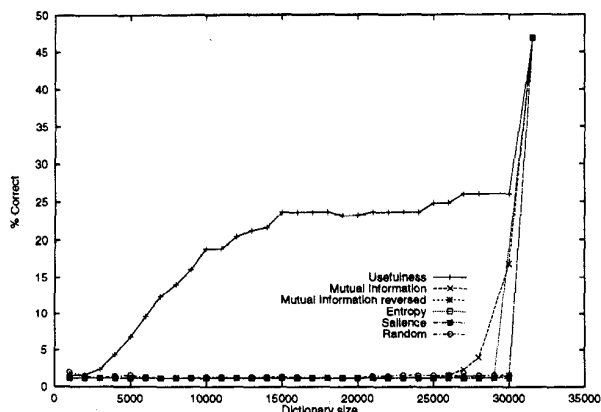


Figure 4: LOB corpus, absolute discounting

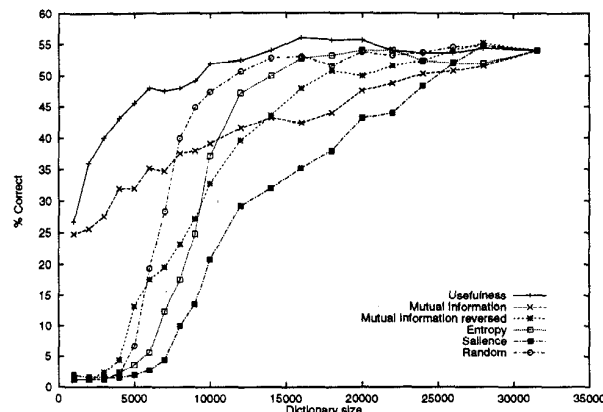


Figure 5: LOB corpus, poisson based

the case of the multinomial, the new usefulness measure bears a striking similarity to the information based measures, no doubt connected with the original derivation of information theory. This resemblance is reflected in the experimental performance: the entropy measure performs better than usefulness for large numbers of keywords.

The behaviour of mutual information is erratic. In particular, the words 'yes' and 'no' corresponding to positive and negative replies in the map task appear as the most useful when ranked by usefulness, but least useful when ranked by mutual information, which produces a word list that is intuitively 'upside down'. The graphs show the effect of simply reversing this list, though with a dubious improvement. In fact, there is no theoretical reason to invert the list. The problems with mutual information are presumably what prompted the invention of salience. Salience, however, still appears from these experiments to perform erratically; sometimes even worse than random. These experiments suggest that entropy would be a better information theoretic measure.

## 6. CONCLUSIONS

The best results in this study have been obtained with the combination of Poisson based probability estimates for words, and the new multi-class usefulness measure. In this case, performance has been shown to improve when the dictionary size is reduced.

It is not clear that there is any theoretically justifiable reason to choose any particular information theoretic measure over another, although experimentally, entropy has been shown to choose good keywords consistently. It is better to derive a measure specifically to maximise discriminability, and in the case of the multinomial, this derivation yields an expression very similar to information theoretic ones.

## 7. REFERENCES

- [1] Michael J. Carey and Eluned S. Parris. Topic spotting using task independent models. In *Proceedings Eurospeech 95, Madrid*, pages 2133–2137, 1995.
- [2] Peter Nowell and Roger K. Moore. The application of dynamic programming techniques to non-word based topic spotting. In *Proceedings Eurospeech '95*, volume 2, pages 1355–1358, Madrid, Spain, September 1995.
- [3] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek. Approaches to topic identification on the switchboard corpus. In *Proceedings ICASSP'94*, volume 1, pages 385–388. IEEE, April 1994.
- [4] Stuart Bird, Sue R. Browning, Roger K. Moore, and Martin J. Russell. Dialogue move recognition using topic spotting techniques. In *Proceedings ESCA Workshop on Spoken Dialogue Systems*, pages 45–48, May 1995.
- [5] Philip N. Garner, Sue R. Browning, Roger K. Moore, and Martin J. Russell. A theory of word frequencies and its application to dialogue move recognition. In *Proceedings ICSLP96*, pages 1880–1883, October 1996.
- [6] Eluned S. Parris and Michael J. Carey. Discriminative phonemes for speaker identification. In *Proceedings ICSLP 94*, volume 4, pages 1843–1846, Yokohama, Japan, September 1994.
- [7] Anne H. Anderson, Miles Bader, Ellen Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jaqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, and Henry S. Thompson. The HCRC map task corpus. *Language and Speech*, 34(4):351–366, 1991.
- [8] Allen L. Gorin. On automated language acquisition. *Journal of the Acoustical Society of America*, 97(6):3441–3461, June 1995.
- [9] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley and sons, Inc., 1968.
- [10] Hermann Ney, Ute Essen, and Reinhard Kneser. On the estimation of 'small' probabilities by leaving-one-out. *IEEE Transactions on pattern analysis and machine intelligence*, 17(12):1202–1212, December 1995.

©British Crown Copyright 1996/DERA Published with the permission of the Controller of Her Britannic Majesty's Stationery Office.



# On topic identification and dialogue move recognition

**Philip N. Garner**

*Speech Research Unit, DERA Malvern, St. Andrews Rd, Malvern,  
Worcestershire WR14 3PS, U.K.*

---

## Abstract

Dialogue move recognition is cited as being representative of a class of problem which may be of concern in data driven natural language processing. The dialogue move recognition problem is formulated as a keyword-based topic identification problem, and is shown to be sensitive to the issue of unknown vocabulary. A model based on the multiple Poisson distribution is shown to alleviate the unknown vocabulary issue, subject to the assumption that the occurrence of keywords represents a small fraction of the data. A keyword selection strategy is derived to ensure this assumption is valid. It is shown that a modified version of Zipf's law provides a suitable prior probability distribution for keywords, and that its inclusion increases classification performance.

© Crown Copyright 1997

---

## 1. Introduction

A Spoken Language Understanding (SLU) system can be thought of as consisting of several different parts; signal processing, a speech recognizer, a language model and finally a natural language or dialogue management module. Generally speaking, current approaches to natural language processing (NLP) tend to be very hand crafted, requiring large amounts of prior knowledge about the structure of language. In stark contrast to this, current speech recognition technology is almost completely data driven. The hypothesis is that SLU technology could be improved by extending the use of data driven methods beyond the speech recognizer into the NLP and dialogue modules.

Several authors have made some progress in this area for specific applications: In the ATIS domain (Cohen, 1995), Schwartz, Miller, Stallard & Makhoul (1996) have developed a model they call a Hidden Understanding Model with the appealing symmetry of modelling higher order semantic features in a similar manner to the way the acoustic features are modelled. Pieraccini and Levin (1995) have developed a system called CHRONUS (Conceptual Hidden Representation Of Natural Unconstrained Speech), which also uses Markovian models to describe semantic meaning. The work of Gorin (1995) is also highly relevant. Several laboratories are also working on data driven dialogue modules for the VERBMOBIL project: Reithinger and Maier (1995)

describe a statistical dialogue model for predicting dialogue events, and Schmitz and Quantz (1996) show that knowledge of dialogue acts is necessary in a translation system.

This work is concerned with methods that may be useful in a data driven SLU scenario, without necessarily defining the scenario. Dialogue act recognition provides a convenient test-bed for such methods. The particular database we use is the HCRC map task corpus (Andersen *et al.*, 1991), which has been annotated at the dialogue move level. Dialogue moves are discussed by Kowtko, Isard & Doherty (1993). The basis of dialogue moves is that when two people engage in a conversation they play a series of games, with constituent moves, in order to impart some piece of information. In the particular case of the map task corpus, 12 distinct moves have been identified; many more are identified in the VERBMOBIL project (Jekat *et al.*, 1995).

Dialogue move recognition involves classification of an input utterance, be it acoustic or text (in this paper only text is considered), into one of  $M$  categories, and in this sense the problem is identical to that of topic identification. In its simplest form, topic identification is a two class problem, where the classes are referred to as “wanted” and “unwanted”. The input can be the word level output of a speech recognizer (Carey & Parris, 1995), or acoustic features (Nowell & Moore, 1995). Recently, with the advent of the Switchboard corpus, the problem has been extended to the multi class domain (e.g. McDonough, Ng, Jeanrenaud, Gish & Rohlicek, 1994).

The purpose of this paper is to formalize the theory used for topic identification in the case of a closed set of  $M$  classes such that it can be applied to dialogue move recognition in a robust manner. The utility of the theory is demonstrated by applying it to the problem of dialogue move recognition on the map task corpus.

## 2. Probabilistic formulation of topic identification

### 2.1. Relationship with language modelling

Given an observation,  $\mathbf{x}$ , typically representing a sequence of words of a particular category, the problem is to infer the category from which it was sampled, also given some labelled training data,  $\mathbf{D}$ . Formally, the category is a sample  $m$  from the set  $\mathcal{M} = \{m_1, m_2, \dots, m_M\}$ , and the solution is to assign  $\mathbf{x}$  to the value of  $m$  resulting from

$$\max_i P(m = m_i | \mathbf{x}, \mathbf{D}).$$

This expression can be “inverted” via Bayes’ theorem to yield

$$P(m = m_i | \mathbf{x}, \mathbf{D}) = \frac{P(\mathbf{x} | m = m_i, \mathbf{D})P(m = m_i | \mathbf{D})}{\sum_{i=1}^M P(\mathbf{x} | m = m_i, \mathbf{D})P(m = m_i | \mathbf{D})} \quad (1)$$

$$\propto P(\mathbf{x} | m = m_i, \mathbf{D})P(m = m_i | \mathbf{D}).$$

Notice that  $P(\mathbf{x} | m = m_i, \mathbf{D})$  is a class dependent language model (LM); this can be made more clear by considering the speech recognition problem. In a speech recognizer, one is presented with an acoustic representation,  $\mathbf{a}$ , of a sequence of words to be recognized (an utterance). A probability,  $P(\mathbf{w} | \mathbf{a}, \mathbf{D})$ , must be attached to a hypothesized sequence,



MATRIX I. Loss matrix in topic identification

	Unwanted	Wanted
Treat as unwanted	$L_{UU}$ (OK)	$L_{WU}$ (False reject)
Treat as wanted	$L_{UW}$ (False accept)	$L_{WW}$ (OK)

$\mathbf{w}$ , of words which could have generated  $\mathbf{a}$  originally. This probability can again be expanded using Bayes' rule:

$$P(\mathbf{w}|\mathbf{a}, \mathbf{D}) \propto P(\mathbf{a}|\mathbf{w}, \mathbf{D})P(\mathbf{w}|\mathbf{D}),$$

the final term being the (class independent) LM. Substituting  $\mathbf{x}$  for  $\mathbf{w}$ , and conditioning the LM term on some class highlights the similarity. Topic identification, then, can be thought of as discriminative language modelling.

### 2.2. The two class case

The two class case is worthy of particular mention as it is traditionally formulated from a decision theoretic point of view: Bayesian decision theory requires utility to be attached to combinations of classifications and actions, that is, define a loss matrix  $\mathbf{L}$  with elements  $L_{ij}$  being some notional loss associated with performing action  $j$  when  $x$  belongs to class  $i$  (see Matrix I). If  $j=W$  denotes "treat as wanted", (for instance, have an operator listen to a report), and  $j=U$  denotes "treat as unwanted" (for instance, ignore the report) then  $L_{WU}$  is the loss associated with treating  $x$  as unwanted when it is actually wanted. For the time being, if  $\mathcal{M}$  is redefined as  $\mathcal{M} = \{W, U\}$ ; the expected loss when assigning  $x$  to class  $W$  is then

$$\begin{aligned} L_W &= L_{WW}P(m=W|\mathbf{x}, \mathbf{D}) + L_{UW}P(m=U|\mathbf{x}, \mathbf{D}) \\ &= L_{WW} \frac{P(\mathbf{x}|W, \mathbf{D})P(W|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} + L_{UW} \frac{P(\mathbf{x}|U, \mathbf{D})P(U|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} \end{aligned}$$

and similarly for  $L_U$ . For readability,  $P(m=W)$  has been abbreviated to  $P(W)$ , and similarly for  $P(U)$ .

To minimize expected loss when there are only two classes, it follows that a decision rule is to classify  $\mathbf{x}$  as  $W$  if and only if

$$\begin{aligned} L_{WW} \frac{P(\mathbf{x}|W, \mathbf{D})P(W|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} + L_{UW} \frac{P(\mathbf{x}|U, \mathbf{D})P(U|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} \\ < L_{WU} \frac{P(\mathbf{x}|W, \mathbf{D})P(W|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})} + L_{UU} \frac{P(\mathbf{x}|U, \mathbf{D})P(U|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})}, \end{aligned}$$

or more simply:

$$(L_{UW} - L_{UU})P(\mathbf{x}|U, \mathbf{D})P(U|\mathbf{D}) < (L_{WU} - L_{WW})P(\mathbf{x}|W, \mathbf{D})P(W|\mathbf{D}).$$

It is generally assumed that the loss due to an incorrect classification is greater than that due to a correct classification, that is  $L_{ij} > L_{ii}$ , in which case the above expression reduces to:

$$\frac{P(\mathbf{x}|W, \mathbf{D})}{P(\mathbf{x}|U, \mathbf{D})} > \frac{P(U|\mathbf{D})}{P(W|\mathbf{D})} \cdot \frac{(L_{UW} - L_{UU})}{(L_{WU} - L_{WW})}. \quad (2)$$

In a real application, the  $L$  terms would be set by someone with some knowledge of the application, and the probabilities on the right hand side (the priors) would be inferred from the data. For evaluation purposes, however, the  $L$  terms are not known and the data is often weighted in favour of the wanted category, so the true prior is unknown; the whole right hand side is generally replaced by a single parameter,  $\lambda$ , which is varied over its range to produce a receiver operating characteristic (ROC) curve.

### 2.3. The multi-class case

The theory in the previous section assumes two classes, and hence can result in a single discrimination metric. Dialogue move recognition is a multi-class problem, and can be thought of as multi-class topic identification. It is tempting to try to use the likelihood ratio as a metric for discrimination of the classes, but alas, for more than two classes, an inequality cannot be formed with a single class on either side.

In the case where one of the classes corresponds to “none of the above”, i.e. a babble topic, topic identification can be formulated as  $M - 1$  two class problems. These can be solved with likelihood ratios and combined into a single ROC curve. Dialogue move recognition however, clearly, corresponds to a “closed set” topic identification problem. Furthermore, in topic identification, one is generally interested in whether the subject is topic or non-topic, and it is correct, and useful, to attach utility at this point. If a car driver wishes to listen to traffic information, it is perfectly reasonable to attach a large loss to missing a report. In dialogue move recognition, however, the dialogue move is not the highest level question in the chain; that might be “Put me through to someone to complain to”, in which case a large loss can be attached to being put through to the wrong telephone extension.

The move recognition can be thought of as being much deeper in the chain, and there is no way a utility can be justified in this problem other than to assign zero loss to a correct classification and equal loss to all possible misclassifications. This is the same as maximizing the likelihood of the move (class). The correct output of the move recognizer is simply a probability for each move, which can be interpreted by the next stage.

Without attaching utility to the various classifications, the correct strategy is to choose the class which maximizes the probability of the class,  $P(m=m_i|\mathbf{x}, \mathbf{D})$ , i.e. to go back to Equation (1).

## 3. Calculation of probabilities

### 3.1. Standard maximum likelihood multinomial approach

Equation (1) requires the calculation of two probabilities: the likelihood of the particular class occurring,  $P(m=m_i|\mathbf{D})$ , and the likelihood that the observation was generated by the model for that class,  $P(\mathbf{x}|m=m_i, \mathbf{D})$ .

The easiest term to calculate is the prior,  $P(m=m_i|\mathbf{D})$ . Note that it is a prior in the sense that it is prior to seeing the observation,  $\mathbf{x}$ ; it is still posterior to the data,  $\mathbf{D}$ . It simply says “What’s the probability that a particular class occurs”. Making the simplification that each class is independent of all previous classes, the intuitive thing to do is to divide the number of times class  $m_i$  occurred in the data by the total number of observations in the data.

The probability,  $P(\mathbf{x}|m=m_i, \mathbf{D})$  is more involved. For the purpose of this paper, assume that the constituent features of  $\mathbf{x}$  are words, generated by repeatedly sampling a variable  $w \in \mathcal{W}$ , where  $\mathcal{W} = \{w_1, w_2, \dots, w_w\}$ . This model is a unigram language model.

The general approach in the literature is to express the likelihood term as the joint probability of the constituent features of  $\mathbf{x}$ .

$$P(\mathbf{x}|m=m_i, \mathbf{D}) = P(w=w_1, w=w_2, \dots, w=w_K|m=m_i, \mathbf{D}), \quad (3)$$

where  $P(w=w_k)$  in this context is taken to mean the probability that  $w$  takes the value of the  $k$ th word in  $\mathbf{x}$ . Given the independence assumption between words, Equation (3) can be expressed as

$$\begin{aligned} P(\mathbf{x}|m_i, \mathbf{D}) &= P(w_1, w_2, \dots, w_K|m=m_i, \mathbf{D}) \\ &= P(w_1|m_i, \mathbf{D})P(w_2|m_i, \mathbf{D}) \cdots P(w_K|m_i, \mathbf{D}), \end{aligned}$$

the notation abbreviated slightly.

Taking “given the move type and the data” to mean “consider only the data that is of that move type”, it is now possible to work out these probabilities. The intuitive method is simply to use the same method as the prior:  $P(w_k|m_i, \mathbf{D})$  can be estimated by taking the number of times that word  $w_k$  occurred in the data of move type  $m_i$ , and dividing by the total number of words in all moves of that type.

### 3.2. An experiment

The HCRC map task corpus (Andersen *et al.*, 1991) has been annotated at the dialogue move level; there are 12 move types in all. The corpus was split into a training and testing set such that no map featured in both sets; in this way, the discrimination could be attributed to the semantic qualities of the text, not the map features.

The training data were used to calculate probabilities as described in the previous section, and these were used to classify the utterances (observations) of the testing data. A confusion matrix is shown in Matrix II.

The horizontal axis represents classification bins, the vertical is the actual class of the utterances. All axes are totalled, so as an example, there were 2459 “Acknowledge” moves in the testing data, 1795 of which were correctly classified. In total 2687 moves were classified as “Acknowledge”.

The classification accuracy is just over 47%; Kowtko *et al.* (1993) state that 70–80% of the moves can be correctly identified by a human (though using context too). The model accuracy is believable given that the model has independence assumptions in the move sequence and in the word sequence.

MATRIX II. An initial confusion matrix for the map task data

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1795	17	32	2	17	66	4	18	119	61	5	323	2459
Align	390	125	19	11	6	33	14	28	114	3	9	8	760
Check	29	38	273	38	46	251	40	40	209	21	37	15	1037
Clarify	7	11	54	35	15	135	7	5	111	8	31	4	423
Explain	23	23	52	15	172	43	11	9	277	82	77	2	786
Instruct	10	30	122	159	35	639	41	20	425	11	49	2	1543
Query-W	5	8	19	4	5	29	186	11	47	1	0	0	315
Query-YN	3	28	47	10	28	36	29	401	144	12	13	3	754
Ready	82	0	4	0	1	11	0	0	9	0	0	0	107
Reply-N	3	1	4	1	1	3	0	1	4	301	3	0	322
Reply-W	11	13	45	20	28	82	10	10	108	21	51	4	403
Reply-Y	329	14	25	4	21	32	3	14	35	11	8	860	1356
Total	2687	308	696	299	375	1360	345	557	1602	532	283	1221	10265

Accuracy = 47.22%

Assuming the test set accuracy is binomially distributed (Bedworth, 1992), the 95% confidence limits for 10 265 independent testing samples are around  $\pm 1\%$ .

The matrix as a whole is reasonably distributed, with large values on the leading diagonal, and smaller values off it. There is a tendency, however, for a lot of utterances to be classified as “Ready”. The “Ready” move is generally played at the start of the conversation, and consists of words like “right” and “okay”. Given that “Acknowledge” also consists of exactly the same words, but is far more frequent, one would expect all the “Ready” moves to actually be classified as “Acknowledge”. It is hypothesized in the next section that this is a symptom of the unknown word or out of vocabulary (OOV) problem.

### 3.3. *The unknown word problem*

When a new utterance is to be classified, a probability must be attached to each word in that utterance. For instance, if the utterance *Go to the left* is to be classified, the probability of each of the words must be evaluated for each of the classes. In a move such as “Instruct”, which is both frequent and has a “rich” language model, most of the words *go*, *to*, *the*, and *left* are likely to have occurred in the training data, and will be given finite probabilities. In a move such as “Clarify”, however, the language model is still rich but the move itself does not occur very often; in such a case, one or more of the words in the utterance to be classified may not have occurred in the training data.

Following the intuitive method, the probability of an unknown word is zero, so the probability of the utterance is zero; the utterance clearly happened, so the model is wrong. In fact, intuition can be updated: it is clear that unknown words will occur, and that their probability ought to be non-zero and will probably be less than that of the least frequent word in that category. The least frequent word that did occur will have occurred once, and a common strategy (e.g. Nowell and Moore, 1995) is to count unknown words as having occurred 0.5 times (some justification for this is hinted at in section 3.4); this is how the confusion matrix in Matrix I was generated.

This ad hoc approach to unknown words explains the bias towards “Ready”: “Ready” is the least frequent move, so the probability attached to an unknown word would be 0.5 divided by some small number being the number of words in that move type in the data. Compare this with a move like “Instruct”, where the unknown word probability is 0.5 divided by a much larger number, due to the rich and frequent nature of that move type. Now imagine that a completely new word occurs in the utterance to be classified: a new map feature, or a nuance of a new talker. The new word will be given a much larger probability by the least frequent class.

### 3.4. *Dice throwing*

The OOV phenomenon is one of the main problems in language modelling, and there is a large amount of literature on the subject. In general, the solution is to apply a statistical smoothing function to the word probabilities, although this can involve a lengthy cross-validation procedure to determine parameters; a recent reference is Ney, Essen and Kneser (1995). The following sections show that for the task of discrimination, a mathematically more attractive approach is available.

When one calculates a probability by dividing the number of occurrences of interest

by the total number of occurrences, one is implicitly assuming a dice throwing model. Statistically, the problem is the same as that of coin tossing or drawing coloured balls from an urn and replacing them. If the number of occurrences of interest is represented by  $n$  and the total number of occurrences by  $N$ , then  $n/N$  is the maximum likelihood (ML) estimate of the true probability. ML estimates traditionally get more accurate as  $n$  and  $N$  get large, and fall over completely as  $n$  tends to zero.

With small databases, especially cases where  $n$  is ever close to zero, it becomes necessary to incorporate prior knowledge in some way. This is traditionally done in classical statistics by assuming some distribution and smoothing the observations to that distribution. In Bayesian statistics, the prior knowledge can be incorporated explicitly, although quantifying prior knowledge is often a problem in itself.

It is instructive to consider the Bayesian formulation of the dice throwing model. Formally, such a model is a multinomial distribution. Appendix A details a Bayesian analysis of this formulation using a flat prior (that is, all combinations of bias are initially estimated to have equal probability), and proves that the result can be applied by replacing the  $n/N$  estimate with

$$\frac{n+1}{N+W},$$

where  $W$  is the total number of possible outcomes (2 for a coin, 6 for a die).

Whilst there is little justification for using a flat prior, this result is useful in that it highlights a fundamental problem: in language modelling,  $W$  is the total vocabulary of the task in question. It can be thought of as the total vocabulary of all the speakers who took part in the task.  $W$  cannot possibly be known; a study by Efron and Thisted (1976), on estimating Shakespeare's vocabulary simply proved that it depends strongly on initial assumptions. The problem has also been tackled by Fisher, Corbet and Williams (1943), Goodman (1949), Good and Toulmin (1956) and McNeil (1973).  $W$  however, is clearly large, and suggests that all probabilities calculated by the simple maximum likelihood model will be grossly overestimated.

Note that in the Bayesian "estimate", the probability when  $n=0$  is half that when  $n=1$ , which justifies in part the  $n=0.5$  estimate in the maximum likelihood case.

### 3.5. The multinomial distribution and topic identification

In fact, the multinomial distribution has other problems when applied to topic identification. Without breaking the sentence down into constituent features, the two class discrimination metric is to classify  $\mathbf{x}$  as wanted if and only if

$$\log \left( \frac{P(\mathbf{x}|W)}{P(\mathbf{x}|U)} \right) > \lambda.$$

Where  $\lambda$  represents the product of the prior ratio and the loss function ratio of Equation (2). The logarithm is generally used for practical convenience. If the words in  $\mathbf{x}$  are considered to be independent,  $P(\mathbf{x})$  can be broken down into the product of the word likelihoods, and the classification rule becomes

$$\sum_{i=1}^K n_i \log \left( \frac{P(w_i|W)}{P(w_i|U)} \right) > \lambda \quad (4)$$

where  $w_i$  represents the  $i$ th word in  $\mathbf{x}$ , and  $\mathbf{x}$  is  $K$  words in length. This equation is often referred to as “accumulated usefulness”.

For the purpose of topic identification, it is clear that only discriminative words need be considered, and these words are termed keywords. The keywords are immediately apparent, being those that result in extreme values of the likelihood ratio. In conventional topic identification, the assumption is made that it is only necessary to compute probabilities for these discriminative words, simply ignoring the others. In fact, it is the probability of the whole utterance that is required.

A more subtle, but very important failure of the multinomial distribution in keyword identification is best demonstrated by example. Consider the problem of spotting weather reports in radio broadcasts. Words that maximize the likelihood ratio are likely to be *rain*, *snow*, *north* and *south*, whilst words that minimize it might be *minister*, *stockmarket* and *Ambridge*. Which class does *the cat sat on the mat* belong in? The purely keyword-based accumulated usefulness equation falls down here, having no evidence whatsoever to make a decision upon. The proper language modelling solution uses default probabilities for unknown words, but these probabilities will be higher for the least frequent class, hence favouring weather forecast given a properly representative database.

What should be acknowledged here is the absence of keywords. The multinomial model correctly applied does this by noticing the presence of other words that are not keywords, but it cannot do this correctly as it does not know the vocabulary. What is needed is a model that explicitly acknowledges zero occurrences of something, whilst ignoring words that it has no knowledge of.

#### 4. Removing the unknown vocabulary problem

##### 4.1. The multiple Poisson distribution

The Poisson distribution was originally derived as an approximation to the binomial distribution. The following brief derivation shows how the multiple Poisson distribution can be derived from the multinomial distribution.

If  $\rho_i$  is defined to be the underlying probability of the event  $w = w_i$ , then the multinomial distribution is

$$P(\mathbf{n}|\mathbf{\rho}) = \frac{N!}{n_1!n_2! \cdots n_w!} \rho_1^{n_1} \rho_2^{n_2} \cdots \rho_w^{n_w}.$$

where  $\mathbf{\rho}$  is the vector  $(\rho_1, \rho_2, \dots, \rho_w)$ , and  $\mathbf{n}$  is the vector  $(n_1, n_2, \dots, n_w)$ . The sum of the components of  $\mathbf{\rho}$  is constrained to be unity.

Making the substitution  $\lambda_i = N\rho_i$ , and replacing  $\rho_w$  with the sum to unity constraint,

$$P(\mathbf{n}|\boldsymbol{\lambda}) = \frac{N!}{n_1!n_2! \cdots n_w!} \left(\frac{\lambda_1}{N}\right)^{n_1} \left(\frac{\lambda_2}{N}\right)^{n_2} \cdots \left(\frac{\lambda_{W-1}}{N}\right)^{n_{W-1}} \\ \times \left(1 - \frac{\lambda_1}{N} - \frac{\lambda_2}{N} - \cdots - \frac{\lambda_{W-1}}{N}\right)^{n_w}.$$

rearranging yields

$$P(\mathbf{n}|\boldsymbol{\lambda}) = \frac{\lambda_1^{n_1} \lambda_2^{n_2} \cdots \lambda_{W-1}^{n_{W-1}}}{n_1! n_2! \cdots n_{W-1}!} \frac{N(N-1)(N-2) \cdots (n_w+1)}{N^{N-n_w}} \\ \times \left(1 - \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_{W-1}}{N}\right)^{n_w}.$$

it can be shown that as  $n \rightarrow \infty$

$$\left(1 - \frac{x}{n}\right)^n \rightarrow e^{-x},$$

so the limiting case turns out to be

$$P(\mathbf{n}|\boldsymbol{\lambda}) = \frac{\lambda_1^{n_1} \lambda_2^{n_2} \cdots \lambda_{W-1}^{n_{W-1}}}{n_1! n_2! \cdots n_{W-1}!} e^{-\lambda_1 - \lambda_2 - \cdots - \lambda_{W-1}},$$

where  $n$  does not contain  $n_w$ , and  $\boldsymbol{\lambda}$  is the obvious thing. Note that  $\lambda_w$  and  $n_w$  have disappeared. This is just the product of  $W-1$  independent Poisson distributions—the multiple Poisson distribution. The approximation is valid if  $N$  is large and  $n_w$  is also large compared to the sum of the other  $n$ .

In fact, the key point here is that  $\rho_w$  does not exist in the Poisson distribution. In the multinomial case there is a certain amount of redundancy in that a  $d$  dimensional multinomial actually has the constraint that all the  $d$  probabilities add to one; it is actually a  $d-1$  dimensional distribution. The redundancy in the  $p$  terms is mirrored in the  $n$  terms, in that if the sum of the  $n$  ( $N$ ) is known, one of the  $n$  is consequently redundant. The Poisson distribution ties down  $N$  to a fixed (infinite) value, so  $n_w$  is redundant. In turn, this is mirrored in the  $\lambda$  terms.

The fact that one term disappears is useful, for example: in a keyword based system, all of the non-keywords can be grouped together and referred to as a single word, the unknown word. If it is this unknown word that is dropped in the above derivation, the result is an expression that is independent of unknown words. Given that all unknown words are grouped together, regardless of how many there are, the result is also vocabulary independent. Re-interpreting the approximations in the derivation, the multiple Poisson distribution is valid for a large training database where the number of occurrences of keywords is small.



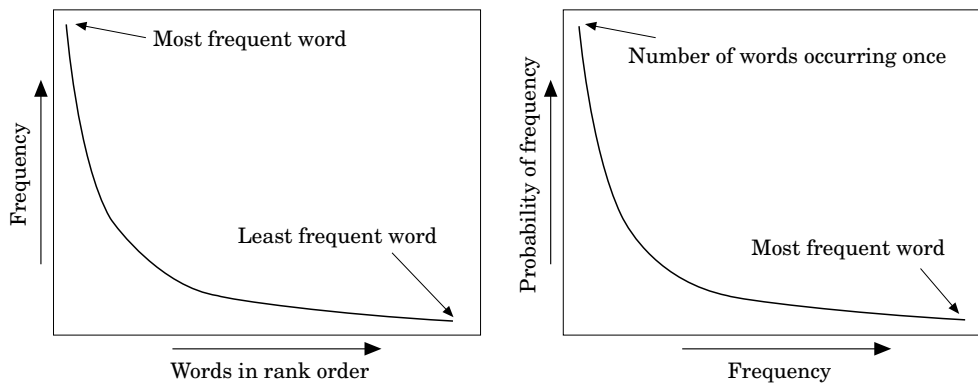


Figure 1. The Zipf plot and how to modify it to relate to probability.

The multiple Poisson is clearly the better method to use if the approximations in its derivation are valid. This distribution has two clear advantages prior to running an experiment:

- (1) The absence of a word has a finite probability, that is, if any or all of the test observation word frequencies are zero then  $P(\mathbf{x}|\mathbf{D})$  is finite. This means the absence of keywords can be penalized.
- (2) There is a default probability for unknown words, that is, if any or all of the training word frequencies are zero then  $P(\mathbf{x}|\mathbf{D})$  is still non-zero.

## 5. Prior information

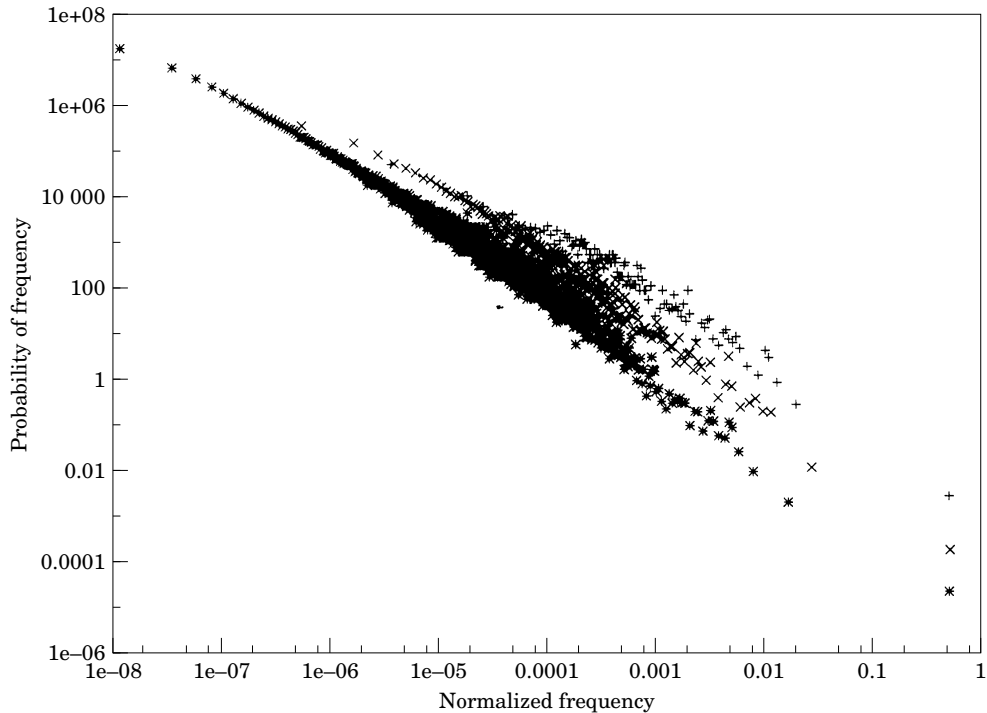
### 5.1. Zipf's law

The parameters of the Poisson distribution are unknown, but information about them is available via the training data. Classically, the training data would be used to estimate the values of these parameters in a maximum likelihood sense. In more recent years the Bayesian approach has found favour, resulting in either integrating out the unknown parameters or calculating a Maximum a posteriori (MAP) estimate. Practically, the two approaches tend to produce similar results in the absence of prior information; in this case though, prior information exists in the form of Zipf's law (Zipf, 1935).

Zipf's law itself is an empirical law relating frequencies of words. If a graph is plotted of frequency as ordinate, and the words rank ordered on the abscissa, that is, the most frequent word on the left and the least frequent on the right, the points will form a smooth curve with approximately reciprocal square root form; the actual analytical form is discussed by McNeil (1973). Further, this law will hold no matter which database is used.

Such a graph is not very useful in that form, but integrating up the vertical axis produces a graph which, suitably normalized, can be interpreted as "Probability of Frequency", which in turn is the prior on the  $\lambda$  terms in the Poisson distribution. This is illustrated in Fig. 1, where the graph on the left is a traditional Zipf plot, and the one on the right is modified as described.

The graph on the right of Fig. 1 can be estimated with a histogram from a large

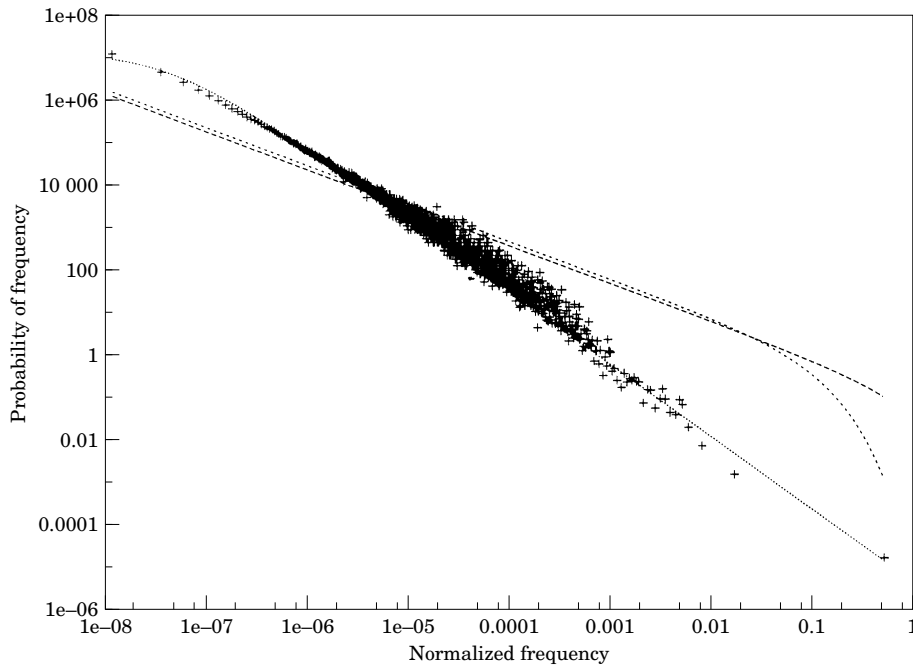


**Figure 2.** Modified Zipf plot for various data sources; Map task (+), King James Bible (x), Wall Street Journal (\*).

dataset, and this is depicted in Fig. 2. The plots refer to the 35 million word ARPA Wall Street Journal corpus, the King James version of the Bible (less than 1 million words), and the entire HCRC map task corpus (less than 200 000 words). Three things are apparent from this plot:

- (1) All the plots are straight lines with (approximately) the same gradient on log-log axes. If the gradients are indeed the same, then Zipf's law holds, and one dataset can be used as a prior for another.
- (2) The smaller data sets have higher tails (the right hand end in this case). This is a well known effect, and suggests that the large dataset is a better approximation to the true distribution.
- (3) The fact that they are straight lines on a double logarithmic scale implies that the real curve is of the form  $y = Ax^m$ , where  $A$  is some normalizing term and  $m$  is the gradient of the line.

Note that the map task plot is only shown for reference. The information in the plots is supposed to be prior information, and looking at any of the map task data is cheating, never mind looking at all of it.



**Figure 3.** Various fits to the Wall Street Journal data; Wall Street Journal (+), Gamma 1 (---), Gamma 2 (-.-), Line Fit (···).

### 5.2. Parameterization of prior information

To be useful as a prior distribution, some convenient parameterized form must be made to fit the Zipf plot. The gamma distribution, defined as

$$P(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda},$$

has an  $x^n$  term, so it ought to be possible to fit a gamma distribution to this database. Fig. 3 illustrates this. The line labelled “Gamma 1” is a gamma distribution with parameters  $\alpha=0.1$  and  $\beta=0$ ; “Gamma 2” is the same with  $\beta=10$ . Shrinking  $\alpha$  any more has the effect of moving the whole line downwards.

There is clearly nothing to be gained from setting  $\beta$  to be anything other than 0; even a value of 10 introduces more curvature than can be justified. Setting  $\alpha$  to some small value may clearly be of benefit though.

A gamma distribution has the advantage of mathematical convenience, being a conjugate prior for a Poisson distribution. Rather than insisting on conjugacy and going out of the way to make a gamma distribution fit the prior information, it ought to be possible to find a distribution that fits the prior information, but is not necessarily conjugate. In Fig. 3, it is clear that the line labelled “Line fit” fits the data much better than the gamma distributions. This is simply the line  $y = Ax^{-1.7}$ , where  $A$  was chosen to make the line go through the data rather than above or below it.

The gamma distribution is not proper (does not integrate to 1) if the  $\alpha$  term falls below  $-1$ , so the line fit is out of the range of the gamma distribution. It is possible, however, to alter  $y \propto x^{-1.7}$  such that it is proper by moving the whole graph to the left by an amount  $\delta$  such that it actually crosses the  $y$  axis. This is equivalent to modelling the prior as

$$P(x) \propto (x + \delta)^{-\gamma},$$

where  $\gamma$  is the (negative) gradient of the line on double logarithmic axes and  $\delta$  is some small number. The value of  $\delta$  can be obtained by evaluating the normalizing constant

$$P(x) = \frac{\gamma - 1}{\delta^{1-\gamma}} (x + \delta)^{-\gamma},$$

and fitting to the histogram. In fact,  $\delta$  controls the general “height” of the line on the graph.

Appendix B shows that assuming a Poisson model, where the parameters follow a gamma distribution, yields the probability of a sequence of words to be

$$P(\mathbf{x}|\mathbf{D}) = \prod_{i=1}^V \frac{\Gamma(x_i + n_i + \alpha)}{\Gamma(n_i + \alpha)} \frac{(D + \beta)^{n_i + \alpha}}{(D + \beta + K)^{x_i + n_i + \alpha}}, \quad (5)$$

where  $\mathbf{x} = \{x_1, x_2, \dots, x_K\}$ ,  $x_k$  is the number of times word  $w_k$  occurred in the observation,  $n_i$  is the number of times word  $w_i$  occurred in the data,  $\mathbf{D}$ , and  $D$  is the total number of words in  $\mathbf{D}$ . In practice, all of these terms are conditioned on the class too. Note that the definition of  $\mathbf{x}$  has been slightly overloaded here to refer to a vector of word counts.

Matrix III shows a confusion matrix for the data with the prior set from line “Gamma 1” in Fig. 3 ( $\alpha=0.1$ ,  $\beta=0$ ). Note that only one move is now categorized as “Ready”, as was the problem with the ML multinomial. There is no category that scoops up all the unclear observations either. As a result, the overall accuracy is higher than that for the ML multinomial.

Appendix C proves that the equivalent of Equation (5) for a “log-linear” prior is

$$P(\mathbf{x}|\mathbf{D}) = \prod_{i=1}^V \frac{(x_i + n_i)!}{n_i!} \frac{U(\gamma, \gamma - x_i - n_i, (D + K)\delta)}{U(\gamma, \gamma - n_i, D\delta)} \frac{D^{1+n_i-\gamma}}{(D + K)^{1+x_i+n_i-\gamma}}, \quad (6)$$

where  $U(a, b, z)$  is Kummer’s confluent hypergeometric function sometimes known as  $\Psi(a; b; z)$ .

Matrix IV shows a confusion matrix for the “log-linear” prior. The classification accuracy is a little less than for a gamma prior, but within the 95% confidence limits. There is a slight bias towards classifying moves as “Ready”, and this is detrimental (more wrongly than correctly classified). On the whole, though, no clear conclusions can be drawn about the relative benefits of the two priors.

In the case of the multinomial, it was clear how to assign a “flat” prior to the distribution by simply setting all the hyperparameters to 1. In this case, however, a flat

MATRIX III. Confusion matrix for measured gamma prior

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1851	25	39	2	37	86	4	23	1	58	9	324	2459
Align	397	171	28	9	24	59	14	38	0	3	9	8	760
Check	41	42	326	28	109	359	28	53	0	11	23	17	1037
Clarify	12	13	69	28	37	212	4	9	0	4	30	5	423
Explain	42	37	101	12	379	86	9	23	0	35	58	4	786
Instruct	21	36	164	74	88	1052	27	34	0	6	39	2	1543
Query-W	9	15	34	3	9	39	187	17	0	0	2	0	315
Query-YN	12	32	70	3	74	81	25	438	0	3	13	3	754
Ready	87	1	4	0	2	12	0	0	0	0	1	0	107
Reply-N	6	1	8	1	10	3	0	1	0	289	3	0	322
Reply-W	22	18	56	16	83	130	6	14	0	9	44	5	403
Reply-Y	343	15	32	2	40	38	3	14	0	3	9	857	1356
Total	2843	406	931	178	892	2157	307	664	1	421	240	1225	10265

Accuracy = 54.77%

MATRIX IV. Confusion matrix for measured "log-linear" prior

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1838	21	41	3	45	91	6	23	5	56	10	320	2459
Align	392	157	27	14	38	58	10	39	1	3	13	8	760
Check	37	41	312	37	121	343	34	56	0	10	30	16	1037
Clarify	12	17	73	32	37	199	5	11	0	4	30	3	423
Explain	42	34	96	14	395	83	6	23	0	29	60	4	786
Instruct	14	36	161	70	111	1043	30	34	3	3	36	2	1543
Query-W	9	16	33	3	9	34	193	16	0	0	2	0	315
Query-YN	11	32	65	4	105	78	26	417	0	1	14	1	754
Ready	82	1	4	0	2	13	0	0	4	0	1	0	107
Reply-N	7	1	11	2	13	2	0	3	0	280	3	0	322
Reply-W	21	19	58	14	90	128	6	12	0	6	45	4	403
Reply-Y	341	16	44	2	43	35	8	11	0	1	10	845	1356
Total	2806	391	925	195	1009	2107	324	645	13	393	254	1203	10265

Accuracy = 54.17%

prior is less clear cut. By inspection, the gamma distribution can be made flat by setting  $\alpha = 1$  and  $\beta = 0$ . This prior essentially says that all the  $\lambda_i$  have an equal probability of lying anywhere from zero to infinity. This is plainly ridiculous; even if all the  $\lambda_i$  were 1, then each observation on average would be expected to contain the entire vocabulary of the task.

Matrix V shows a confusion matrix for classification using a “flat” gamma prior and probabilities calculated using Equation (5). Considering the prior, the results are remarkably good.

A flat prior is a mathematical convenience, though. A prior should either be non-informative or represent real prior information. The next section shows that the addition of Zipf’s law can be even more beneficial when training data is scarce.

## 6. Evaluation

In order to evaluate the different methods of assigning probabilities to observations of sequences of words, classification experiments were performed on various amounts of training data. Of the original 64 dialogues in the training set, 10 were used as a “burn in” set to ensure that at least some data from each move type was present. Classifications were then performed, adding another dialogue to the training data each time.

The hypothesis was that the approaches using prior information should perform better than those without for small amounts of training data. In addition, the use of a log-linear prior should improve on the conjugate gamma prior.

Figure 4 shows the classification performance for the various methods for the range of training data sizes used. The behaviour is broadly as predicted: all of the Poisson based measures outperform the standard multinomial, and the inclusion of prior information increases performance for small amounts of training data.

The log-linear prior does not perform as well as expected, though. In fact, the gamma prior is consistently better. The reason for this is most likely to be that the log-linear fit is only a somewhat *ad hoc* attempt to fit the Zipf plot. Whilst it fits the visible part of the plot, there is no reason to believe that it fits the unseen part to the extreme left. In fact, the log-linear curve bends downwards to cross the axis in this region, and the unseen plot is unlikely to do this. In turn, it is this region which is most important from the point of view of reverting to prior information because it contains the unknown words. The gamma distribution has two advantages here: it does not actually cross the axis, and for larger amounts of data it does not dictate a particular functional form, i.e., the functional form with a gamma prior is the same as for a flat prior.

The evaluation as shown is somewhat biased in that certain moves (notably “Acknowledge”) are very easy to classify, and are very prevalent. A more objective evaluation should use a test set with equal probability of occurrence of any particular move. This is reflected in Fig. 5: A test set was constructed by randomly sampling 100 observations of each type of move from the original test set, and this knowledge was reflected by ignoring  $P(m_i|\mathbf{D})$ . The effect of this is to increase “performance resolution”. The overall performance is lower reflecting the lower frequency of easy to classify moves, but the curves are now separated, emphasizing the importance of prior information. The 95% confidence limits on the classification rate for this smaller test set are around  $\pm 3\%$ .

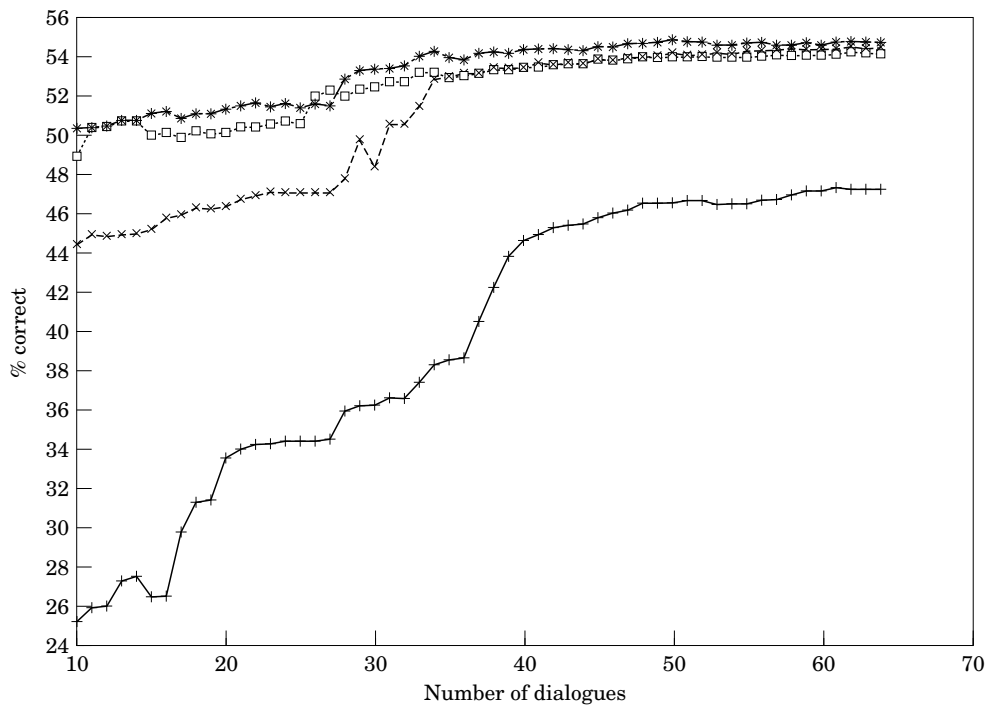
In the latter figure, the curves for the two informative priors are coincident for a while, but separate when there is a large amount of data, although they still lie within each other’s 95% confidence limits. It can be concluded at this stage that there is

MATRIX V. Confusion matrix for flat gamma prior

	AGE	AGN	CCK	CFY	EIN	ICT	Q-W	QYN	RDY	R-N	R-W	R-Y	Total
Acknowledge	1916	15	36	0	36	107	3	22	0	4	3	317	2459
Align	404	159	23	2	23	93	8	39	0	0	1	8	760
Check	55	24	303	6	101	464	12	50	0	1	6	15	1037
Clarify	15	10	73	7	32	265	1	9	0	0	6	5	423
Explain	43	18	117	2	430	134	2	28	0	1	8	3	786
Instruct	23	20	117	17	82	1233	10	30	0	0	10	1	1543
Query-W	11	19	46	0	11	50	157	21	0	0	0	0	315
Query-YN	11	25	80	0	59	118	9	447	0	0	4	1	754
Ready	88	1	4	0	1	13	0	0	0	0	0	0	107
Reply-N	188	1	10	0	35	8	0	3	0	75	2	0	322
Reply-W	19	9	51	6	105	181	2	11	0	0	16	3	403
Reply-Y	352	8	30	0	41	50	0	14	0	2	4	855	1356
Total	3125	309	890	40	956	2716	204	674	0	83	60	1208	10265

Accuracy = 54.53%





**Figure 4.** Classification performance vs. amount of training data for the four different probability measures. ML multinomial (+), Poisson, flat prior (x), Poisson, gamma prior (\*), Poisson, log-linear prior (□).

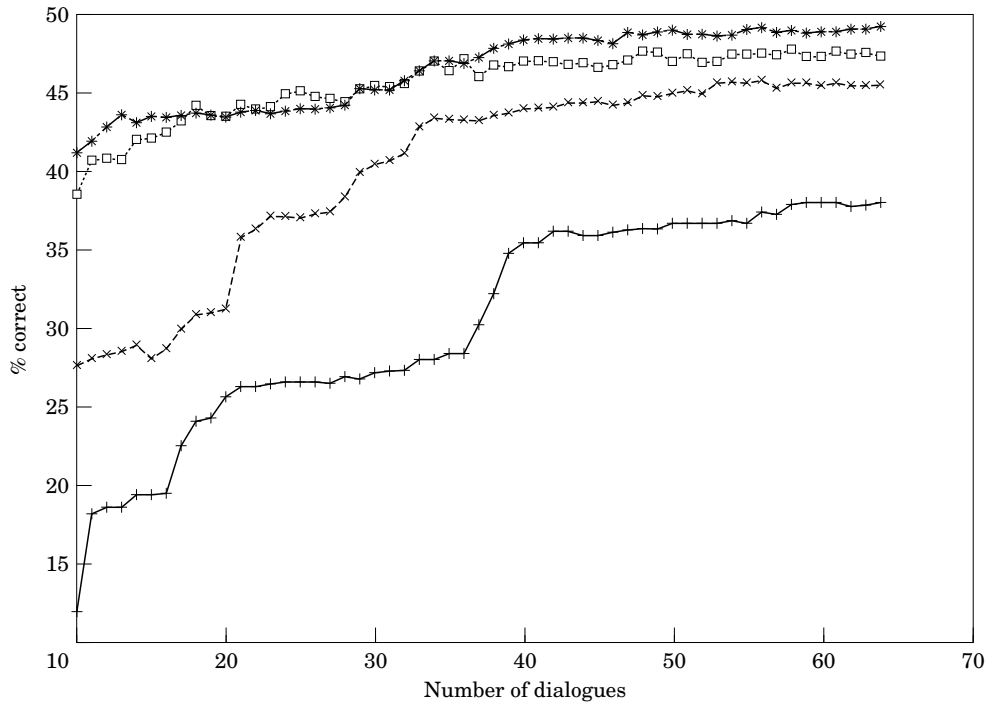
nothing to be gained from using the log-linear prior, especially since the functional form is unnecessarily complicated.

## 7. Pruning the vocabulary—keyword identification

### 7.1. Discussion

In the preceding sections, no attempt has been made to choose those words that are discriminative. The vocabulary size has been defined as the complete vocabulary of the training data. In fact, it is clear that some words will have a much greater discriminative effect than others, and even that some words will have no discriminative effect at all. Further, the multiple Poisson approximation to the multinomial becomes more valid as the combined rate of occurrence of vocabulary words decreases. Pruning the vocabulary should therefore increase the performance of the model. One can imagine some optimal vocabulary that is small enough to allow the Poisson approximation to be valid, yet large enough to retain discriminability.

In the traditional topic identification scenario, the discriminative words are chosen as those that maximize the ratio (4), and are referred to as keywords. This ratio is referred to as usefulness because it identifies those words that are useful. In the multi-



**Figure 5.** Classification performance vs. amount of training data for the four different probability measures using a test set with equal move distribution; ML multinomial (+), Poisson, flat prior (x), Poisson, gamma prior (\*), Poisson, log-linear prior (□).

class case, however, the single ratio does not apply and it is less clear how to attach a discriminability measure to words.

### 7.2 A multi-class discriminability measure

The decision rule itself can be used to indicate the measure of discriminability for each word in the vocabulary: the overall decision rule is to maximize

$$P(m_i|\mathbf{x}, \mathbf{D}) = \frac{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})}{P(\mathbf{x}|\mathbf{D})}$$

$$= \frac{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})}{\sum_{j=1}^M P(\mathbf{x}|m_j, \mathbf{D})P(m_j|\mathbf{D})}$$

over all moves in  $\mathcal{M}$ . This is the same as minimizing the reciprocal, in which case the summation appears in the numerator and the expression breaks down into a sum of ratios:

$$\mathcal{P}_i = \frac{P(\mathbf{x}|m_1, \mathbf{D})P(m_1|\mathbf{D})}{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})} + \frac{P(\mathbf{x}|m_2, \mathbf{D})P(m_2|\mathbf{D})}{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})} \\ + \cdots + \frac{P(\mathbf{x}|m_M, \mathbf{D})P(m_M|\mathbf{D})}{P(\mathbf{x}|m_i, \mathbf{D})P(m_i|\mathbf{D})},$$

which consists of easily differentiable parts.

When choosing a feature set, it is desirable to choose features that have maximum effect upon the decision rule. Consider an observation,  $\mathbf{x}$ , consisting of a single word  $w_k$ . The difference in  $\mathcal{P}_i$  after observing  $\mathbf{x}$  is likely to be proportional to

$$\left. \frac{\partial \mathcal{P}_i}{\partial x_k} \right|_{x_k=0}.$$

where  $x_k$  is the number of times words  $w_k$  occurred in  $\mathbf{x}$ . It is natural to use the expectation of this expression over all words in the vocabulary:

$$E\left(\frac{\partial \mathcal{P}_i}{\partial x_k}\right) = \sum_{k=1}^V \frac{\partial \mathcal{P}_i}{\partial x_k} P(w_k|m_i, \mathbf{D}),$$

and since the problem is multi-class, an expectation can also be taken over classes.

$$E\left(\frac{\partial \mathcal{P}}{\partial x_k}\right) = \sum_{i=1}^M E\left(\frac{\partial \mathcal{P}_i}{\partial x_k}\right) P(m_i|\mathbf{D}).$$

Interchanging the order of summation, the contribution of a particular word  $w_k$  to this expression is

$$U(w_k) = \sum_{i=1}^M \frac{\partial \mathcal{P}_i}{\partial x_k} P(w_k|m_i, \mathbf{D}) P(m_i|\mathbf{D}).$$

It follows that, since  $\mathcal{P}_i$  is to be minimized for a correct classification, words should be chosen which minimize  $U(w_k)$ .

Consider first the case where the words are assumed to be distributed multinomially. The probability of a sequence of words  $\mathbf{x}$  conditioned on the class, in a maximum likelihood sense, is

$$P(\mathbf{x}|m_i) = \prod_{k=1}^K \frac{n_{ik}}{D_i},$$

where, with a change of notation to allow conditioning on the class, there are  $n_{ik}$  words of type  $w_k$  and  $D_i$  words in total in class  $m_i$  of the training set. Differentiating as prescribed and setting  $x_k=0$  results in

$$U(w_k) = \sum_{i=1}^M \frac{n_{ik}}{D_i} \frac{n_i}{N} \sum_{\substack{j=1 \\ j \neq i}}^M \frac{n_j}{n_i} \log \frac{n_{jk} D_i}{n_{ik} D_j},$$

where there are  $n_j$  examples of class  $m_j$  in the training data. In practice, the two  $n_i$  terms cancel, and the  $N$  is unnecessary.

In the special case of two classes, this expression can be written

$$\begin{aligned} U(w_k) = & -P(m_2)P(w_k|m_1) \log \frac{P(w_k|m_1)}{P(w_k|m_2)} \\ & -P(m_1)P(w_k|m_2) \log \frac{P(w_k|m_2)}{P(w_k|m_1)}. \end{aligned}$$

Each of these terms is exactly the same as that given by Parris and Carey (1994), though from a much more general view, and corresponds to combining features indicative of the wanted class with features indicative of the unwanted class. For this reason, the name usefulness is retained. Curiously though, the term corresponding to class 1 is weighted by the probability of class 2 and vice-versa.

In the case of the Poisson based estimate for word probability, consider one of the terms of  $\mathcal{P}_i$ , comparing move  $j$  with move  $i$ :

$$\frac{\prod_{k=1}^V \left[ \frac{\Gamma(n_{jk} + \alpha + x_k)}{\Gamma(n_{jk} + \alpha)} \frac{(D_j + \beta)^{n_{jk} + \alpha}}{(D_j + \beta + K)^{n_{jk} + \alpha + x_k}} \right] \frac{n_j}{N}}{\prod_{k=1}^V \left[ \frac{\Gamma(n_{ik} + \alpha + x_k)}{\Gamma(n_{ik} + \alpha)} \frac{(D_i + \beta)^{n_{ik} + \alpha}}{(D_i + \beta + K)^{n_{ik} + \alpha + x_k}} \right] \frac{n_i}{N}},$$

rearranging yields

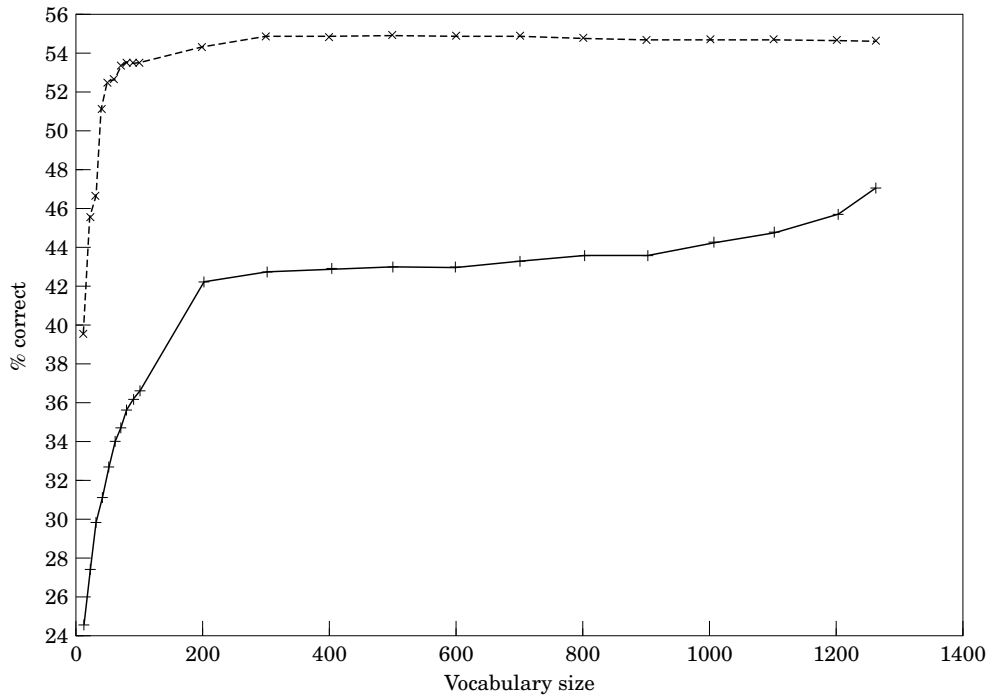
$$\prod_{k=1}^V \left[ \frac{\Gamma(n_{jk} + \alpha + x_k)}{\Gamma(n_{ik} + \alpha + x_k)} \frac{\Gamma(n_{ik} + \alpha)}{\Gamma(n_{jk} + \alpha)} \frac{(D_j + \beta)^{n_{jk} + \alpha} (D_i + \beta + K)^{n_{ik} + \alpha + x_k}}{(D_i + \beta)^{n_{ik} + \alpha} (D_j + \beta + K)^{n_{jk} + \alpha + x_k}} \right] \frac{n_j}{n_i}.$$

Differentiating with respect to a single  $x_k$  yields the same expression multiplied by

$$\log(D_i + \beta + K) - \log(D_j + \beta + K) + \psi(n_{jk} + \alpha + x_k) - \psi(n_{ik} + \alpha + x_k),$$

where  $\psi$  is the digamma function. Setting all the  $x_k = 0$  as before, and  $K = 0$ , the expression for the usefulness of word  $w_k$  becomes

$$\begin{aligned} U(w_k) = & \sum_{i=1}^M P(w_k|m_i) \sum_{\substack{j=1 \\ j \neq i}}^M n_j [\log(D_i + \beta) - \log(D_j + \beta) \\ & + \psi(n_{jk} + \alpha) - \psi(n_{ik} + \alpha)], \end{aligned}$$



**Figure 6.** The effect on the classification rate of pruning the vocabulary; ML multinomial (+), Poisson, gamma prior (x).

where  $P(w_k|m_i)$  is the probability of an observation consisting of the single word  $w_k$ .

The usefulness in the Poisson case is essentially the same form as that for the multinomial (the two logarithm terms can be written as the logarithm of a ratio), with the addition of the digamma functions. Digamma functions simply relate gamma functions to their first derivatives. Practically, the expression is more complicated to compute as the  $P(w_k|m_i)$  term is a product over all keywords, but mathematically the result is reassuringly simple.

### 7.3. Evaluation

The 64 dialogues of the same training set as before were used to generate ordered lists of words for the ML multinomial and Poisson with gamma prior probability measures. Classification experiments were then performed using all 64 training dialogues and the same test situations as before, but with various vocabulary sizes. The results for the full test set are shown in Fig. 6.

Figure 7 shows the same results, but for the equally distributed test set used before. The features of the Poisson based curve are enhanced in the latter figure.

There is a definite peak in the Poisson curve at 300 keywords which corresponds to an optimal vocabulary size. To the right of this point, the performance of the multinomial continues to increase as the unknown vocabulary becomes less of a problem. The performance of the Poisson based system deteriorates though. One reason for this is

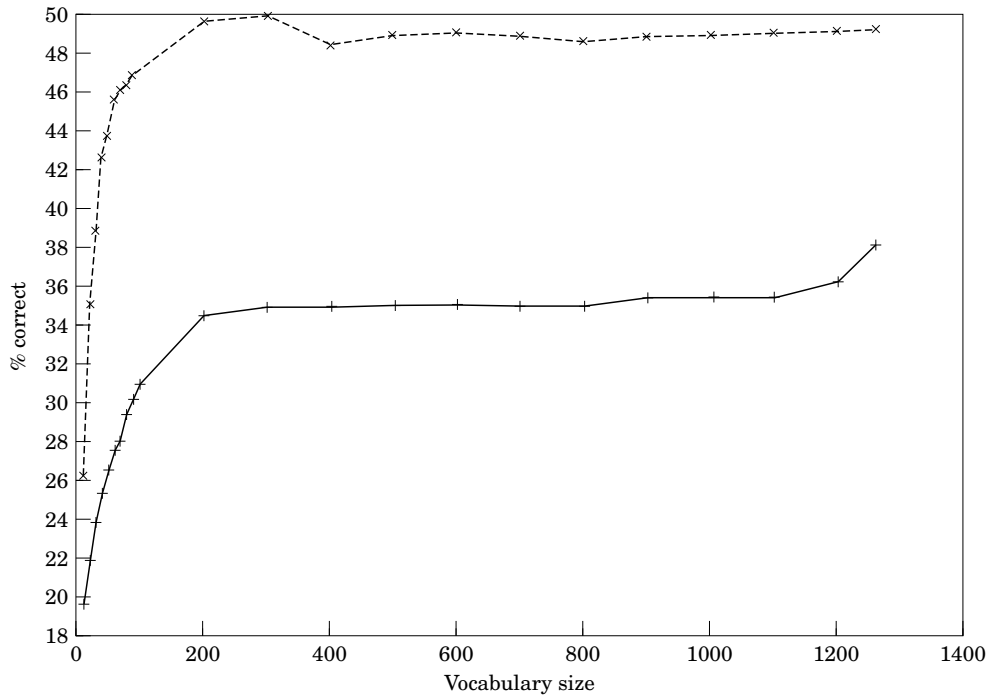


Figure 7. The effect on the classification rate of pruning the vocabulary for equal move probability; ML multinomial (+), Poisson, gamma prior (x).

clearly the failing nature of the approximation in the derivation of the multiple Poisson distribution.

The other reason is that the system is including words which only occurred once in the training set and are not discriminative. If a word only appears once, it will be treated as positively discriminative for the move in which it appears. With fewer than 300 keywords both systems deteriorate, and with fewer than 100 words there is simply not enough information to retain performance.

Results were reported by Garner and Hemsworth (1997), comparing several other methods of pruning the vocabulary. These results are summarized in Fig. 8, which shows the effect of pruning the vocabulary for a Poisson based model, using various pruning strategies. The key labels refer to measures as follows: usefulness is discussed in this paper, and the line is identical to that in Fig. 6. Mutual information,

$$I(\mathcal{M}; w_k) = \sum_{i=1}^M \log \frac{P(m_i | w_k)}{P(m_i)} P(m_i),$$

is the information provided by word  $w_k$  about the set of moves  $\mathcal{M}$ . Mutual information reversed can be thought of as  $-I(\mathcal{M}; w_k)$ , and was used because it was not clear whether to maximize large positive or negative values. Entropy is defined as

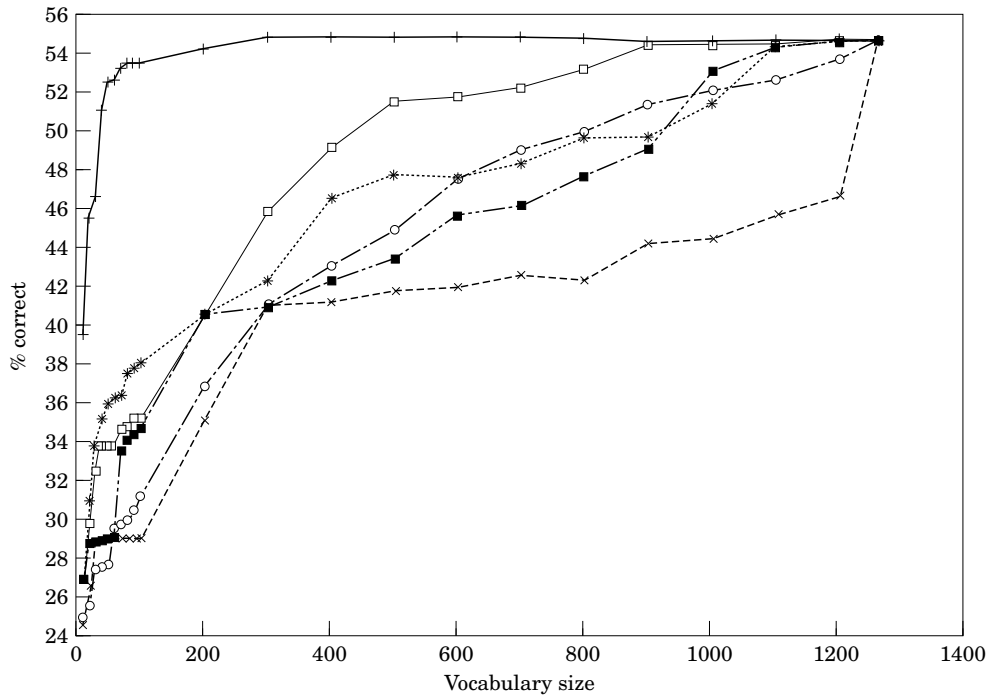


Figure 8. Comparison of various vocabulary pruning methods for the Poisson based model; Usefulness (+), Mutual information (×), Mutual information reversed (\*), Entropy (□), Saliency (■), Random (○).

$$I_E(\mathcal{M}; w_k) = - \sum_{i=1}^M P(m_i) \log P(m_i) + \sum_{i=1}^M P(m_i|w_k) \log P(m_i|w_k),$$

and represents the increase in entropy of the ensemble  $\mathcal{M}$  when word  $w_k$  is observed. Saliency is defined as

$$S(\mathcal{M}; w_k) = \sum_{i=1}^M P(m_i|w_k) \log \frac{P(m_i|w_k)}{P(m_i)},$$

and is used by Gorin (1995) in his language acquisition work. The line labelled "Random" is simply a random pruning of the vocabulary. It is clear that usefulness outperforms all other methods considered in this experiment.

### 8. Conclusions

This paper has outlined a consistent and rigorous approach to keyword-based topic identification, resulting in a robust enough theory to give good results when applied to dialogue move recognition. This leads to the practical result that dialogue moves can be inferred using a unigram language model to an accuracy of around 50%. The

approach, however, is more important than the actual result. Dialogue move recognition can clearly be much improved using dialogue context and acoustic intonation: Reithinger, Engel, Kipp and Klesen (1996) report an accuracy of around 40% predicting 18 intentional dialogue acts from a VERBMOBIL corpus using purely dialogue context, and Taylor, Shimodaira, Isard and Kowtko (1996) report approximately 55% on the map task corpus using purely intonation. In a more genuine experiment on a similar corpus again using intonation (Taylor, King, Isard, Wright and Kowtko, 1997), a move accuracy of around 39% is reported, which goes up to around 44% when dialogue history is considered.

In short, it is suggested that the multiple Poisson distribution is a better distribution with which to model words than a multinomial, since it alleviates the unknown vocabulary problem. This advantage far outweighs the approximate nature of the distribution. When the approximation is taken into account and only discriminative words are chosen, the multiple Poisson distribution performs even better.

Zipf's law provides a convenient subjective linguistic prior to incorporate into the posterior probability in a Bayesian sense. Its inclusion further improves performance.

This paper only goes as far as suggesting that there is some optimal vocabulary to use for a particular task; it does not suggest how to find that vocabulary, other than the obvious use of a validation set.

An assumption taken throughout is that the word and dialogue move boundaries are known, which is not the case in the context of, for instance, automatic speech recognition (ASR). Any extension to ASR would need to acknowledge the uncertain nature of the transcription, and one possible approach would be the use of lattices. The probability of a single utterance could then be evaluated as the sum of the probabilities of the words in each path through a lattice, weighted by the probability of the path. This, however, remains a subject for future research. The problem of detection of move boundaries has been addressed by Cettolo and Corazza (1997).

In addition to thanking the anonymous referees for their time and comments, I should like to extend my gratitude to the following colleagues for their help at various stages of this work. From DERA: Sue Browning, Martin Russell, Roger Moore, Mark Bedworth, Anthony Brown, Wendy Holmes and Richard Glendinning, from Forum Technology: Jörg Ueberla, and from University of Edinburgh: Steve Isard, Jacqueline Kowtko and Cathy Sotillo. We acknowledge HCRC for the use of the dialogue game annotations for the map task corpus.

### References

- Anderson, A. H., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C. & Thompson, H. S. (1991). The HCRC map task corpus. *Language and Speech* **34**(4), 351–366.
- Bedworth, M. D. (1992). On the quality and quantity of data and pattern recognition. Memorandum, Defence Research Agency, St Andrews Rd, Malvern, WORCS, WR14 3PS, UK.
- Carey, M. J. & Parris, E. S. (1995). Topic spotting using task independent models. In *Proceedings Eurospeech 1995, Madrid*, pp. 2133–2137.
- Cettolo, M. & Corazza, A. (1997). Automatic detection of semantic boundaries. In *Proceedings Eurospeech '97*, **5**, Rhodes, pp. 919–922.
- Cohen, J. R. (ed.) (1995). *Proceedings of the Spoken Language Systems Technology Workshop, Barton Creek Resort Conference Center, Austin, Texas*. ARPA, Morgan Kaufmann Publishers, Inc., San Francisco.
- Efron, B. & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**(3), 435–47.
- Fisher, R. A., Corbet, A. S. & Williams, C. B. (1943). The relation between the number of species and the



- number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, **12**, 42–58.
- Garner, P. N. & Hemsforth, A. (1997). A keyword selection strategy for dialogue move recognition and multi-class topic identification. In *Proceedings ICASSP 1997*. IEEE, **3**, 1823–1826.
- Good, I. J. & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, **43**, 45–63.
- Goodman, L. A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics*, **20**, 572–579.
- Gorin, A. L. (1995). On automated language acquisition. *Journal of the Acoustical Society of America*, **97**(6), 3441–3461.
- Gradshteyn, I. S. & Ryzhik, I. M. (1980). *Table of Integrals, Series, and Products*. Academic Press, 5th Edn.
- Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M. & Quantz, J. J. (1995). Dialogue acts in VERBMOBIL. VERBMOBIL report 65.
- Kowtko, J. C., Isard, S. D. & Doherty, G. M. (1993). Conversational games within dialogue. Technical report, Human Communication Research Centre, University of Edinburgh, 2 Buccleugh Place, Edinburgh EH8 9LW SCOTLAND.
- McDonough, J., Ng, K., Jeanrenaud, P., Gish, H. & Rohlicek, J. R. (1994). Approaches to topic identification on the switchboard corpus. In *Proceedings ICASSP 1994*, volume **1**, pp. 385–388. IEEE.
- McNeil, D. R. (1973). Estimating an author's vocabulary. *Journal of the American Statistical Association*, **68**(341), 92–96.
- Ney, H., Essen, U. & Kneser, R. (1995). On the estimation of 'small' probabilities by leaving-one-out. *IEEE transactions on pattern analysis and machine intelligence*, **17**(12), 1202–1212.
- Nowell, P. & Moore, R. K. (1995). The application of dynamic programming techniques to non-word based topic spotting. In *Proceedings Eurospeech 1995*, volume **2**, pp. 1355–1358, Madrid, Spain.
- O'Hagan, A. (1994). *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*. Edward Arnold.
- Parris, E. S. & Carey, M. J. (1994). Discriminative phonemes for speaker identification. In *Proceedings ICSLP 1994*, volume **4**, pp. 1843–1846, Yokohama, Japan.
- Pieraccini, R. & Levin, E. (1995). A spontaneous-speech understanding system for database query applications. In *Proceedings ESCA Workshop on Spoken Dialogue Systems*, pp. 85–88.
- Reithinger, N. & Maier, E. (1995). Utilizing statistical dialogue act processing in VERBMOBIL. VERBMOBIL report 80, Reprint from *ACL-95 Proceedings*.
- Reithinger, N., Engel, R., Kipp, M. & Klesen, M. (1996). Predicting dialogue acts for a speech-to-speech translation system. VERBMOBIL report 151, also in *Proceedings of ICSLP 1996*, pp. 654–657.
- Schmitz, B. & Quantz, J. J. (1996). Dialogue acts in automatic dialogue interpreting. VERBMOBIL report 173, also in *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, TMI-95, Leuven, pp. 33–47.
- Schwartz, R., Miller, S., Stallard, D. & Makhoul, J. (1996). Language understanding using hidden understanding models. In *Proceedings ICSLP 1996*, pp. 997–1000.
- Taylor, P., Shimodaira, H., Isard, S., King, S. & Kowtko, J. (1996). Using prosodic information to constrain language models for spoken dialogue. In *Proceedings ICSLP 1996*, volume **1**, pp. 216–219.
- Taylor, P., King, S., Isard, S., Wright, H. & Kowtko, J. (1997). Using intonation to constrain language models in speech recognition. In *Proceedings Eurospeech 1997*, **5**, Rhodes, pp. 2763–2768.
- Zipf, G. K. (1935). *The Psycho-Biology of Language*. Houghton-Mifflin, Boston.

(Received 20 May 1997 and accepted for publication 11 August 1997)

#### Appendix A: probability “estimates” from the multinomial

The following proof is by no means new, but is included in an abbreviated form for reference. For a more complete discussion of the techniques involved, see any text on Bayesian statistics.

A predictive distribution,  $P(\mathbf{x}|m_i, D)$ , is sought, where  $\mathbf{x}$  is a sequence of words. To simplify the notation, assume that all of the calculations in this section are conditioned on  $m = m_i$ , i.e.  $P(\mathbf{x}|\mathbf{D}) \equiv P(\mathbf{x}|m = m_i, \mathbf{D})$ . Assume that  $\mathbf{x}$  was generated by repeatedly sampling a variable  $w$  from the set  $\mathcal{W} = \{w_1, w_2, \dots, w_w\}$ . If there are  $K$  words in  $\mathbf{x}$ ,

$$P(\mathbf{x}|\mathbf{D}) = P(w = w_1, w = w_2, \dots, w = w_k|\mathbf{D}).$$

If there are  $n_k$  words of type  $k$  in  $\mathbf{D}$ , and  $N$  words altogether, and the unconditional probability  $P(w = w_k)$  of each word is  $\rho_k$ , the core problem is to find

$$P(\mathbf{x}|\mathbf{D}) = P(\mathbf{x}|\mathbf{n}, N) = \int_0^1 \cdots \int_0^1 d\boldsymbol{\rho} P(\mathbf{x}|\boldsymbol{\rho}, \mathbf{n}, N) P(\boldsymbol{\rho}|\mathbf{n}, N),$$

where  $\boldsymbol{\rho} = \{\rho_1, \rho_2, \dots, \rho_W\}$ , and the notation is meant to mean “integrate w.r.t. each  $\rho_k$ ”.

With reference to, for instance O’Hagan (1994), applying Bayes’ theorem to the final term and assuming  $\boldsymbol{\rho}$  follows a Dirichlet distribution, if there are  $x_i$  words of type  $w_i$  in  $\mathbf{x}$ , then

$$\begin{aligned} P(\mathbf{x}|\mathbf{D}) &= \int_0^1 \cdots \int_0^1 d\boldsymbol{\rho} \rho_1^{x_1} \rho_2^{x_2} \cdots \rho_W^{x_W} \frac{\rho_1^{n_1 + \alpha_1 - 1} \rho_2^{n_2 + \alpha_2 - 1} \cdots \rho_W^{n_W + \alpha_W - 1}}{\iint d\boldsymbol{\rho} \rho_1^{n_1 + \alpha_1 - 1} \rho_2^{n_2 + \alpha_2 - 1} \cdots \rho_W^{n_W + \alpha_W - 1}} \\ &= \frac{B(n_1 + \alpha_1 + x_1, n_2 + \alpha_2 + x_2, \dots, n_W + \alpha_W + x_W)}{B(n_1 + \alpha_1, n_2 + \alpha_2, \dots, n_W + \alpha_W)}, \end{aligned}$$

where  $B(a, b, \dots)$  is the multivariate beta function.

It is actually more informative to look at this equation for a specific sequence: the terms for any word that does not appear in  $\mathbf{x}$  simply cancel, leaving terms for the words that do occur, so the probability of the sequence  $\{w_1, w_1, w_2, w_2\}$  is

$$\frac{n_1 + 1}{N + W} \frac{n_1 + 2}{N + W + 1} \frac{n_2 + 1}{N + W + 2} \frac{n_2 + 2}{N + W + 3}.$$

A flat prior has been assumed by setting  $\alpha_1 = \alpha_2 = \alpha_W = 1$ . Notice that the expression is equivalent to adding each word of the observation,  $\mathbf{x}$ , to the data,  $\mathbf{D}$ , before evaluating the next word. This effect is sometimes known as Laplace’s rule.

### Appendix B: word probabilities from the multiple Poisson

Given an observation  $\mathbf{x}$  containing  $x_k$  words of type  $w_k$ , the probability  $P(\mathbf{x}|\mathbf{D})$  is required. This depends upon the parameters of the Poisson distribution and is given by

$$P(\mathbf{x}|\mathbf{D}) = \int_0^\infty \cdots \int_0^\infty d\boldsymbol{\lambda} P(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{D}) P(\boldsymbol{\lambda}|\mathbf{D}). \quad (\text{B.1})$$

This integral is actually a lot easier than it looks because the multiple Poisson distribution is simply the product of  $V$  independent Poisson distributions.

An important consideration here is that of “window size”. In the case of the multinomial, the probability of generating  $n$  sets of  $l$  words is the same as the probability

of generating a single set of  $n \times l$  words. For the multiple Poisson, the concept of window size is more important, and the two cases are different. One approach would be to choose a window size natural to the application such as length of observation. Observations are generally of different length, though, so the approach taken here is to normalize the window to be one word long, and to treat an observation length  $l$  as  $l$  separate observations.

Consider the univariate version of  $P(\lambda|\mathbf{D})$ : the univariate Poisson distribution is defined to be

$$P(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}.$$

If a sequence of  $D$  trials results in observations  $\mathbf{n} = \{n_1, n_2, \dots, n_D\}$ , then

$$\begin{aligned} P(\mathbf{n}|\lambda) &= \frac{\lambda^{n_1} e^{-\lambda}}{n_1!} \frac{\lambda^{n_2} e^{-\lambda}}{n_2!} \dots \frac{\lambda^{n_D} e^{-\lambda}}{n_D!} \\ &= \frac{\lambda^{n_1+n_2+\dots+n_D} e^{-D\lambda}}{n_1! n_2! \dots n_D!} \\ &= \lambda^n e^{-D\lambda}. \end{aligned}$$

The  $n_k$  can be either 1 or 0 corresponding to a word either appearing or not appearing, whereas the  $n$  refers to the number of occurrences in the  $D$  trials. Using Bayes' theorem to obtain the posterior,

$$P(\lambda|\mathbf{D}) = \frac{P(\mathbf{n}|\lambda)P(\lambda)}{\int_0^{\infty} d\lambda P(\mathbf{n}|\lambda)P(\lambda)}.$$

Assuming that  $P(\lambda)$  is a gamma distribution, and that the normalizing terms cancel,

$$\begin{aligned} P(\lambda|\mathbf{D}) &= \frac{\lambda^n e^{-D\lambda} \lambda^{\alpha-1} e^{-\beta\lambda}}{\int_0^{\infty} d\lambda \lambda^n e^{-D\lambda} \lambda^{\alpha-1} e^{-\beta\lambda}} \\ &= \frac{\lambda^{n+\alpha-1} e^{-(D+\beta)\lambda}}{\int_0^{\infty} d\lambda \lambda^{n+\alpha-1} e^{-(D+\beta)\lambda}} \\ &= \frac{(D+\beta)^{n+\alpha}}{\Gamma(n+\alpha)} \lambda^{n+\alpha-1} e^{-(D+\beta)\lambda}. \end{aligned}$$

Assuming that all the  $\lambda_i$  are drawn from the same gamma distribution, the multivariate case is simply the product of these over all words, that is

$$P(\boldsymbol{\lambda}|\mathbf{D}) = \prod_{i=1}^V \frac{(D+\beta)^{n_i+\alpha}}{\Gamma(n_i+\alpha)} \lambda_i^{n_i+\alpha-1} e^{-(D+\beta)\lambda_i}.$$

The likelihood term of (B.1) is simply the raw multivariate Poisson distribution,

$$P(\mathbf{x}|\boldsymbol{\lambda}, \mathbf{D}) = P(\mathbf{x}|\boldsymbol{\lambda}) = \prod_{k=1}^K \left( \prod_{i=1}^V \frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right),$$

where  $x_{ik}$  is the number of words of type  $w_i$  in position  $k$  in  $\mathbf{x}$ ,  $\mathbf{x}$  being  $K$  words in length.

Equation (B.1) can be rearranged and evaluated as  $V$  independent integrals thus:

$$\begin{aligned} P(\mathbf{x}|\mathbf{D}) &= \int_0^\infty \cdots \int_0^\infty d\boldsymbol{\lambda} \prod_{k=1}^K \left( \prod_{i=1}^V \frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right) \prod_{i=1}^V \frac{(D+\beta)^{n_i+\alpha}}{\Gamma(n_i+\alpha)} \lambda_i^{n_i+\alpha-1} e^{-(D+\beta)\lambda_i} \\ &= \prod_{i=1}^V \left[ \frac{(D+\beta)^{n_i+\alpha}}{\Gamma(n_i+\alpha)} \int_0^\infty d\lambda_i \prod_{k=1}^K \left( \frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right) \lambda_i^{n_i+\alpha-1} e^{-(D+\beta)\lambda_i} \right] \\ &= \prod_{i=1}^V \left[ \frac{(D+\beta)^{n_i+\alpha}}{\Gamma(n_i+\alpha)} \prod_{k=1}^K \frac{1}{x_{ik}!} \int_0^\infty d\lambda_i \lambda_i^{x_i+n_i+\alpha-1} e^{-(D+K+\beta)\lambda_i} \right] \end{aligned}$$

Since  $x_{ik}$  is always either 1 or 0, and the integral is now just another gamma integral, the final form is

$$P(\mathbf{x}|\mathbf{D}) = \prod_{i=1}^V \left[ \frac{\Gamma(x_i+n_i+\alpha)}{\Gamma(n_i+\alpha)} \frac{(D+\beta)^{n_i+\alpha}}{(D+\beta+K)^{x_i+n_i+\alpha}} \right].$$

In practice, this equation simplifies in that if  $V \gg K$ ,  $x_i$  will mostly be zero and the gamma functions cancel. Further,  $\prod_V (\cdot)^{n_i+\alpha} = (\cdot)^{N+V\alpha}$ , so the product is only over  $K$  terms.

This distribution is related to the negative binomial distribution. Consider for the moment the terms inside the product, which can be written

$$\frac{(x_i+n_i+\alpha-1)!}{x_i!(n_i+\alpha-1)!} (1-p)^{n_i+\alpha} p^{x_i}$$

where  $p = (D+\beta-1)^{-1}$ . The  $x_i$  disappeared in the derivation since it was always 0 or 1. This expression is of the form

$$P(x|r, p) = \binom{r+x-1}{x} p^r q^x$$

which is the negative binomial distribution that Fisher used to count butterflies (Fisher, Corbet & Williams, 1943) and that Efron & Thisted (1976) used to model Shakespeare's output. The derivation differs from Fisher in its use of a prior distribution, and since the  $\alpha$  terms are not necessarily integer, the normalizing term cannot be written using factorials.

### Appendix C: Poisson distribution with "log-linear" prior

Following the notation and argument in Appendix B,  $P(\lambda|\mathbf{D})$  is required. This is the product of all the univariate cases, where a single univariate case is given by

$$P(\lambda|n) = \frac{\lambda^N e^{-D\lambda} (\lambda + \delta)^{-\gamma}}{\int_0^{\infty} d\lambda \lambda^N e^{-D\lambda} (\lambda + \delta)^{-\gamma}},$$

assuming the normalizing constants cancel.

The integral in the denominator can be solved by noticing the similarity with the integral definition of the confluent hypergeometric function (Gradshteyn & Ryzhik, 1980):

$$\Gamma(a)U(a, b, z) = \int_0^{\infty} e^{-zt} t^{a-1} (1+t)^{b-a-1} dt.$$

Making the change of variable  $t = \lambda/\delta$ , the integral in the denominator becomes

$$\begin{aligned} I &= \int_0^{\infty} dt \delta(\delta t)^N e^{-D\delta t} (\delta + \delta t)^{-\gamma} \\ &= \delta^{N+1-\gamma} \int_0^{\infty} dt t^N e^{-Dt} (1+t)^{-\gamma} \\ &= \delta^{N+1-\gamma} \Gamma(N+1) U(N+1, N+2-\gamma, D\delta). \end{aligned}$$

Changing notation to allow for the multivariate case, an proceeding as in Appendix B,

$$\begin{aligned}
P(\mathbf{x}|\mathbf{D}) &= \int_0^\infty \cdots \int_0^\infty d\lambda \prod_{k=1}^K \left( \prod_{i=1}^V \frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right) \\
&\quad \times \prod_{i=1}^V \frac{\lambda_i^{n_i} e^{-D\lambda_i} (\lambda_i + \delta)^{-\gamma}}{\delta^{n_i+1-\gamma} \Gamma(n_i+1) U(n_i+1, n_i+2-\gamma, D\delta)} \\
&= \prod_{i=1}^V \left[ \frac{\int_0^\infty d\lambda_i \prod_{k=1}^K \left( \frac{\lambda_i^{x_{ik}} e^{-\lambda_i}}{x_{ik}!} \right) \lambda_i^{n_i} e^{-D\lambda_i} (\lambda_i + \delta)^{-\gamma}}{\delta^{n_i+1-\gamma} \Gamma(n_i+1) U(n_i+1, n_i+2-\gamma, D\delta)} \right] \\
&= \prod_{i=1}^V \left[ \frac{\prod_{k=1}^K \frac{1}{x_{ik}!} \int_0^\infty d\lambda_i \lambda_i^{x_i+x_k} e^{-(D+K)\lambda_i} (\lambda_i + \delta)^{-\gamma}}{\delta^{n_i+1-\gamma} \Gamma(n_i+1) U(n_i+1, n_i+2-\gamma, D\delta)} \right].
\end{aligned}$$

Again,  $x_{ik}$  can only ever be 0 or 1. The whole expression can be simplified using the Kummer transformation

$$U(a, b, z) = z^{1-b} U(1+a-b, 2-b, z).$$

In addition, some of the  $\delta$  terms cancel, and the arguments to the gamma functions are always integer so factorials can be used, yielding

$$P(\mathbf{x}|\mathbf{D}) = \prod_{i=1}^V \frac{(x_i+n_i)!}{n_i!} \frac{U[\gamma, \gamma-x_i-n_i, (D+K)\delta]}{U(\gamma, \gamma-n_i, D\delta)} \frac{D^{1+n_i-\gamma}}{(D+K)^{1+x_i+n_i-\gamma}}.$$

The relationship with the gamma prior expression is now evident; this expression is like that for a flat prior, but with  $\gamma$  as a notional ‘‘initial count’’ for  $n_i$ , and the addition of the ratio of confluent hypergeometric functions.

## USING FORMANT FREQUENCIES IN SPEECH RECOGNITION

*John N. Holmes (1), Wendy J. Holmes (2) and Philip N. Garner (2)*

(1) Speech Technology Consultant, 19 Maylands Drive, Uxbridge, UB8 1BH, U.K.

Tel: +44 1895 236328, E-mail: jnh@jnholmes.demon.co.uk

(2) Speech Research Unit, DRA Malvern, St. Andrews Road, Malvern, Worcs., WR14 3PS, U.K.

Tel: +44 1684 894104/894157, E-mail: holmes/garner@signal.dra.hmg.gb

### ABSTRACT

Formant frequencies have rarely been used as acoustic features for speech recognition, in spite of their phonetic significance. For some speech sounds one or more of the formants may be so badly defined that it is not useful to attempt a frequency measurement. Also, it is often difficult to decide which formant labels to attach to particular spectral peaks. This paper describes a new method of formant analysis which includes techniques to overcome both of the above difficulties. Using the same data and HMM model structure, results are compared between a recognizer using conventional cepstrum features and one using three formant frequencies, combined with fewer cepstrum features to represent general spectral trends. For the same total number of features, results show that including formant features can offer increased accuracy over using cepstrum features only.

### 1. INTRODUCTION

It has been known for many years that formant frequencies are important in determining the phonetic content of speech sounds. Several authors have therefore investigated formant frequencies as speech recognition features, using various methods for basic analysis, such as linear prediction [1], [2], analysis by synthesis with Fourier spectra [3], and peak picking on cepstrally smoothed spectra [4]. However, using formants for recognition can sometimes cause problems, and they have not yet been widely adopted. It is obvious, for example, that formant frequencies cannot discriminate between speech sounds for which the main differences are unrelated to formants. Thus they are unable to distinguish between speech and silence or between vowels and weak fricatives. Whenever any formants are poorly defined in the signal (e.g. in fricatives), measurements will be unreliable, and it is therefore essential that their estimated frequencies should be given little weight in the recognition process.

To be useful as features for automatic speech recognition, formant frequencies must be supplemented by signal level and general spectral shape information, such as provided by low-order cepstrum features, for example. However, whenever the speech spectrum has a peaky structure, the phonetic detail is better described by formant frequencies than by the more usual higher-order cepstrum features, which have no simple relationship with formant frequencies.

It is impossible to determine from the spectrum of some speech sounds whether a particular peak should be associated with one formant or with a pair, and sometimes a formant may be so weak as a consequence of weak excitation that it causes no peak in the spectrum. Either of

these situations can cause all higher-frequency formants to be wrongly labelled, with disastrous effects on the recognition. In such cases alternative labellings must be produced, and any uncertainties that cannot be resolved in other ways must be resolved within the recognition algorithm. The decisions are thus delayed until the words have been recognized [1]. However, many labelling uncertainties of single frames can be safely resolved merely by applying formant continuity constraints [2], which are a general property of speech. First applying continuity constraints is actually better for the standard HMM formalism, which does not exploit continuity of features.

This paper presents a new method of formant analysis which has provision for dealing with ambiguous labelling and with indistinct formants. The method has been used to supplement low-order cepstrum features for speech recognition.

### 2. NEW METHOD FOR FORMANT ANALYSIS

#### 2.1 Human interpretation of formants

When supplied with a wide-band spectrogram of a speech signal, an expert in experimental phonetics can usually estimate fairly well where the formant trajectories are for all parts of the signal for which such an interpretation would be useful. For those parts of the signal where the formant peaks of a particular spectral cross-section are not well defined, an expert can normally still make a reasonable interpretation by using phonetic knowledge about the normal properties of speech sounds and by interpolation between neighbouring sounds for which the formant structure is clearer. It is generally more difficult to estimate formant frequencies automatically, given the same short-term spectral analysis that is the basis of spectrographic display. However, the task is easy if the spectral cross-section of the signal has a small number of clearly defined peaks. Provided that each of the three lowest-frequency peaks is in the frequency range typical of one of the three lowest formants, only one sensible formant interpretation of the spectral shape is possible.

Fig. 1 shows a spectral cross-section which has clear peaks, with the positions of the formants marked. On these occasions a single spectral cross section is all that is required to make a reliable estimate. Sometimes, however, two formants may be so close in frequency that they give rise to only a single spectral peak. There can also often be occasions where a total of three spectral peaks are visible, but the frequencies and intensities might be such that the middle peak could plausibly be F2 by itself and the third peak be F3, or the middle peak could be F2 and F3 together, with the third peak being F4. In this case even a human expert would be incapable of making a reliable choice,

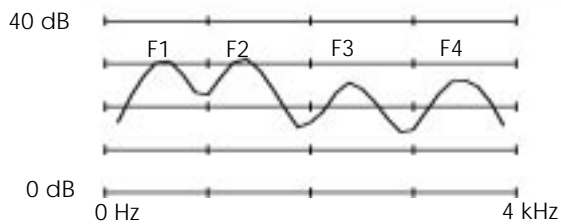


Fig. 1. Spectrum with clear formants

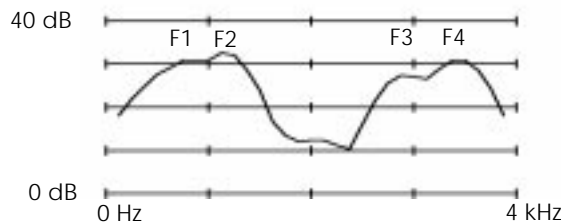


Fig. 2. F1 and F2 in a single spectral peak

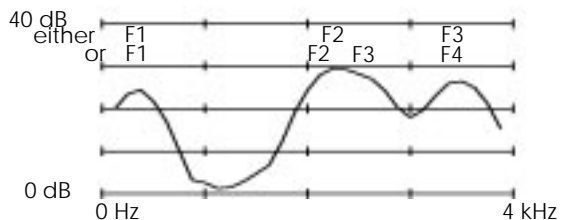


Fig. 3. Ambiguous formant labelling

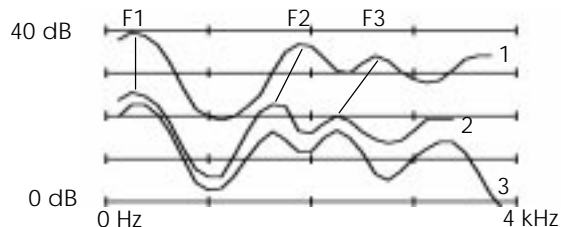


Fig. 4. Frequency warping of pattern (trace 1) into warped pattern (trace 2) to align with input (trace 3)

given only a single spectral cross-section. However, the expert would be able to postulate a small number of plausible alternatives, where in most cases all but one of these alternatives could subsequently be rejected by using continuity constraints. Thus unambiguous formant trajectories would be obtained for a substantial proportion of any utterance. Fig. 2 shows a spectral cross-section for which F1 and F2 are obviously both associated with the lowest-frequency peak, whereas the spectrum shown in Fig. 3 is an example where there is uncertainty about the correct formant labelling, and both of the marked formant allocations would be plausible.

An important novelty of the formant estimation method described in this paper is that it exploits this human ability to apply formant labels to spectral cross-sections, giving alternative formant allocations to peaks where appropriate.

## 2.2 Preliminary formant estimates

The formant analysis uses log power spectra derived from 64-point FFTs of a signal sampled at 8 kHz. To ensure that the cross-sections represent the formants as well as possible, the FFTs are taken from regions immediately after points of excitation of the vocal tract, selected on the basis of a local power maximum. There is a store of about 150 typical spectral cross-sections, each of which is associated with one or more sets of plausible labellings of the lowest three formants, provided by a human expert. Each input spectral cross-section is first compared with all the stored patterns, to select a few which have the most similar general spectral shape. These few patterns are then compared with the input using a dynamic programming (DP) technique in the frequency domain to find the frequency scale warping of the stored patterns which gives the best match to the input. Fig. 4 illustrates a typical warping operation. The DP cost function includes components dependent on spectral level, spectral slope and extent of frequency warping. The pattern with the best DP score and any close competitors are selected for further consideration. The frequency warping of each such pattern is applied to the formant frequencies

stored with the pattern, to give preliminary formant frequency estimates. These estimates are quantized at the 125 Hz spacing of the FFT, and more finely quantized formant frequencies are derived by matching typical formant shapes to the spectrum in the region of the chosen FFT points.

## 2.3 Selection of smooth formant tracks

Any alternative formant labellings given by the few best-fitting patterns are used as input to an additional DP process, which finds the best smooth trajectories through the available formant frequency candidates. A second pass of the DP smoothing process is then made, in which the best formant labelling given by the first pass is used as an additional input to the DP cost function. This second pass will give an alternative smooth path through the available formant candidates if the score for such a path is not much worse than the score of the best path.

The formant analysis method usually gives a unique formant interpretation of speech signals, and never gives more than two different interpretations. Whenever it is apparent from a spectrogram where the formants should be, it is extremely rare for the algorithm to fail to give the correct values, and they are nearly always provided by the first choice. For each output formant frequency an estimate of confidence in the measurement is derived based on spectral level and spectral curvature, so that less reliable formant frequencies can be given less weight in recognition decisions.

## 2.4 Analysis example

Fig. 5 shows a typical spectrogram with superimposed formant tracks. During the [j] and the [t] burst F1 has been omitted because there was no confidence in its accuracy. The two alternative interpretations of F2 and F3 are both reasonable, but the first choice obviously provides correct continuity into the nearby phones. Neither F2 nor F3 could be usefully estimated during the [d] closure, and F2 in the [n] was only given any confidence for one frame. The first choice is clearly correct during the first part of the [eI]



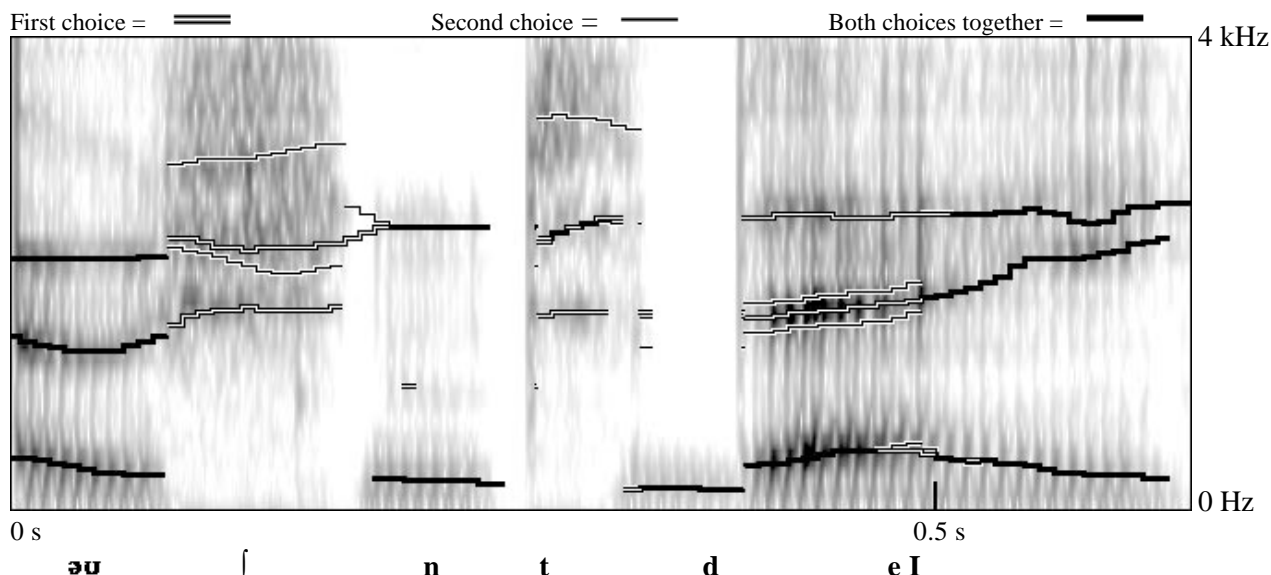


Fig. 5. Spectrogram of the words "ocean today", with superimposed formant tracks. Tracks are not plotted when there is no confidence in their accuracy.

diphthong, but the second choice was initially a plausible interpretation, until the later part of the diphthong had been analysed to reveal the first-choice F2 moving close to F3.

### 3. USING CONFIDENCE ESTIMATES AND AMBIGUITY IN RECOGNITION

Alternative formant sets arising from labelling ambiguity have so far been accommodated in recognition just by choosing the formant set which gives the highest HMM emission probability for each frame and model state.

During silence or background noise, and whenever there is no obvious spectral peak near to the estimated formant frequency, there will be no confidence in the formant frequency estimate, which should not then be used at all in the recognition. In this case, the appropriate formant information to use in the recognizer should be specified by prior information about its likely position. During peaky vowel spectra on the other hand, the measured frequencies will be given high confidence, although there may be occasional labelling ambiguity. There is a continuum of possibilities between these two extremes that can most suitably be accommodated by regarding the uncertainty of formant position as the variance of a notional Gaussian distribution of the true frequency about the estimated value.

The probabilistic interpretation leads naturally to the incorporation of prior knowledge about formant positions when the confidence is low. This prior knowledge is used by shifting the mean of the formant distribution away from the measured value, towards some suitable prior value for that formant. A heuristic procedure has been devised for using the estimated confidence computed from the spectrum to derive a formant measurement standard deviation and bias towards a prior distribution, both expressed in Hz. Although this process is *ad hoc*, it has been found to give plausible values and experimentation has shown that the precise values are not critical to recognition performance.

Assuming that variances are associated with all formant measurements, the HMM emission probability calculation

needs to be modified to allow for a continuum of possible variance values for each formant. It can be shown that in the case of Gaussian models this modification corresponds to a convolution of the formant and model distributions, so that the variances simply add. The use of variance thus provides a sound theoretical framework to represent confidence associated with formant estimates, which is an improvement over an earlier version [5] of the formant-based recognizer, whereby the confidence was simply used as a weight to multiply log probabilities.

### 4. EXPERIMENTS

The aim was to compare recognition results using formant features for describing fine spectral detail with those obtained using a more conventional mel-cepstrum representation. In order to directly assess the usefulness of the formants, the same total number of features was used for both representations, and exactly the same low-order cepstrum features were used for describing general spectral shape. Thus the only difference was in the use of formants versus higher cepstral coefficients for representing detailed spectrum shape. The experiments were performed for the simple task of connected-digit recognition. While the details of the front-end processing and the modelling task have not been optimized to maximize performance, the system provides a good basis for comparative experiments.

#### 4.1 Experimental set-up

The test data were four lists of 50 digit triples spoken by each of 10 male speakers. The training data were from 225 different male speakers, each reading 19 four-digit strings taken from a vocabulary of 10 strings. The output of the FFT was used both to estimate formant frequencies with associated confidence measures and to compute the mel-cepstrum. Experiments were then carried out to compare a representation using the first eight cepstrum coefficients and an overall energy feature, with a feature set in which cepstrum coefficients 6, 7 and 8 were replaced by the three formant features. To provide a basis for comparison, an experiment was also carried out using a representation

Experimental condition	% Correct	% Subs.	% Del.	% Ins.	% Error
5 cepstrum features + energy	95.5	3.5	1.0	0.3	4.8
8 cepstrum features + energy	96.0	3.0	1.0	0.3	4.3
5 cepstrum features + energy + 3 formants	94.0	4.8	1.2	11.6	17.6
Include confidence measure with formants	96.9	2.3	0.8	0.3	3.4
Also include second choice formants	97.1	2.2	0.7	0.3	3.2

Table 1. Connected-digit recognition performance for front-end representations using only cepstrum features compared with a representation with the higher-order cepstral coefficients replaced by formant features.

which simply omitted cepstrum coefficients 6, 7 and 8, so using a total of only six features.

In all cases, three-state context-independent monophone models and four single-state non-speech models were used, all with single-Gaussian pdfs and diagonal covariance matrices. The model structure was a simple left-to-right one which included self-loop transitions. Model means were initialized from a very small quantity of hand-annotated training data (twelve digits from each of two speakers), with all model variances initialized to the same arbitrary value. All model parameters were trained with ten iterations of Baum-Welch re-estimation. During training, an appropriate lower limit was imposed on all the model variance parameters, to prevent them training to unrealistically low values which could prevent generalisation to the test data.

#### 4.2 Treatment of formant features

As a pre-processing stage for both training and recognition, each observed formant value was moved towards its prior by an amount determined by the observation's confidence measure. The result of this stage was that high-confidence formant values were unchanged but, as the confidence decreased, the formant was moved further towards its prior. When there was no confidence, the prior value was used.

The main benefit of the confidence measure and multiple formant hypotheses was expected to be in the recognition stage, as the training process is much more constrained. Therefore, in training, the second choice formant values have not yet been used and no further use has so far been made of the confidence measure. Both were optionally included in the recognition phase, as described in Section 3.

#### 4.3 Results and discussion

The results given in Table 1 show that, provided the degree of reliability in the formant estimation is taken into account, recognition performance is better when using formant features than when using only mel-cepstrum features. When compared with the results using just six cepstrum features, the benefit from adding the three formant features is three times greater than that obtained by adding the three additional cepstrum features.

When alternative formant sets were also included, there was a further small improvement in performance. Only a small improvement was expected because the first-choice values given by this algorithm are usually the correct ones. When they are correct, allowing the second choice could only increase recognition errors. It is therefore clearly desirable to find some way of using an estimate of the relative probabilities of correctness of the first and second choice in the recognition, and this will be included in future research.

The recognition results demonstrate the importance of using formant measurement accuracy in order to obtain good recognition performance. When the formant features were not given special treatment, there were significant problems with insertion errors. These errors were caused by mismatches between the formant frequencies in the non-speech models with those measured for the non-speech regions of the test data. A simple word-insertion penalty did not reduce these errors, but they disappeared when the formant confidence measure was incorporated.

## 5. CONCLUSIONS

These simple experiments have already demonstrated that a recognition system using formant features can provide better performance than one using mel-cepstrum features alone, for the same total number of features. We now need to confirm that similar benefits are obtained on a more challenging task with a larger database. The next stage of algorithm development is to incorporate both the variance representing confidence in formant measurement and the multiple formant hypotheses in an extended Baum-Welch re-estimation process. It is also possible to incorporate the shift of uncertain formant measurements towards their priors within the probabilistic formalism itself, in place of the heuristic approach used here.

Other issues to investigate include the use of time derivative features, which ought to be more valuable for smoothly-changing formants than for high order cepstrum features, particularly because formant transitions are known to be important cues for place of articulation of consonants.

## 6. REFERENCES

- [1] M.J. Hunt, "Delayed Decisions in Speech Recognition - The Case of Formants", Pattern Recognition Letters, Vol. 6, pp. 121-137, July 1987.
- [2] P. Schmid and E. Barnard, "Robust, N-Best Formant Tracking", Proc. EUROSPEECH'95, pp. 737-740, Madrid, 1995.
- [3] L. Welling and H. Ney, "A Model for Efficient Formant Estimation", Proc. IEEE ICASSP, pp. 797-800, Atlanta, 1996.
- [4] Y. Laprie and M.-O. Berger, "Active Models for Regularizing Formant Trajectories", Proc. ICSLP, pp. 815-818, Banff, 1992.
- [5] J.N. Holmes and W.J. Holmes, "The Use of Formants as Acoustic Features for Automatic Speech Recognition", Proc. IOA, Vol. 18, part 9, pp. 275-282, Nov. 1996.

# ON THE ROBUST INCORPORATION OF FORMANT FEATURES INTO HIDDEN MARKOV MODELS FOR AUTOMATIC SPEECH RECOGNITION

*Philip N. Garner, Wendy J. Holmes*

Defence Evaluation and Research Agency,  
St. Andrews Rd, Malvern, WORCS. WR14 3PS, UK

## ABSTRACT

A formant analyser is interpreted probabilistically via a noisy channel model. This leads to a robust method of incorporating formant features into hidden Markov models for automatic speech recognition. Recognition equations follow trivially, and Baum-Welch style re-estimation equations are derived. Experimental results are presented which provide empirical proof of convergence, and demonstrate the effectiveness of the technique in achieving recognition performance advantages by including formant features rather than only using cepstrum features.

## 1. INTRODUCTION

Formant frequencies are known to be important in determining the phonetic content of speech sounds. Formants, however, are not generally used as features for automatic speech recognition as they may be ambiguous or badly defined and do not provide the necessary information for making certain distinctions (such as identifying silence). A new method of formant analysis has recently been presented [1] which includes techniques to overcome the difficulties normally associated with extracting and using formant information. Firstly, in cases of ambiguity, alternative sets of formant frequencies are offered to the recognition process. Secondly, a novel feature of the new formant analyser is that each formant frequency estimate is assigned a measure of confidence. The confidence measure is important because it allows for cases where formants are poorly defined in the signal (e.g. fricatives) so that any single estimate of frequency is likely to be unreliable. In such cases, it is essential that the estimated frequencies are given little weight in the recognition process, and that the recognition decision is based on signal level and general spectral shape information.

Whilst it is clear that the confidence measures have implications when the formants are used as features in speech recognition, it is not obvious how to include such measures in, for instance, an HMM based system. In this paper, we present a method for interpreting confidence estimates which can then be rigorously incorporated into a probabilistic model.

## 2. INTERPRETATION OF THE CONFIDENCE MEASURE

The formant analyser produces a confidence value for each formant for each time frame. This value represents and

estimate of the confidence in the accuracy of the formant frequency measurement, and is derived automatically based on spectral level and curvature. The confidence values are represented as standard deviations which, when squared, can be thought of as variances of normal distributions centred upon the formant estimates. Interpreted in this way, the formant analyser emits the parameters of a normal distribution representing its belief about the position of each formant. When the confidence is high, the variance is low, representing strong belief in the estimate, and weak belief outside it. At the other extreme, a low confidence represents a high variance representing almost equal belief in all possible frequencies. This belief oriented interpretation is necessarily Bayesian.

## 3. MATHEMATICAL FORMULATION

### 3.1. Recognition

In conventional hidden Markov modelling, a state is assumed to emit an observation,  $\mathbf{y}_t$ , according to some output distribution. In this paper, we will assume that the output probability distribution for state  $j$  is a single multivariate normal with mean  $\boldsymbol{\mu}_j$  and covariance matrix  $\boldsymbol{\Sigma}_j$ . The required probability at time  $t$  is

$$\Pr(\mathbf{y}_t | \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}) = \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}).$$

With the formant analyser, the observation comprises both a formant vector,  $\mathbf{f}_t$ , and a confidence vector,  $\mathbf{c}_t$ . The actual feature vector, being the real values of the formant frequencies, is unknown. The confidence measure of the formant analyser is assumed here to take the form of variance.

Given that we observe a distribution, the required expression for the output probability of the state is now

$$\Pr(\mathbf{f}_t, \mathbf{C}_t | \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}),$$

where  $\mathbf{C}_t$  is the (diagonal) matrix of formant variances. The most informative way to proceed is to expand this expression thus

$$\Pr(\mathbf{f}_t, \mathbf{C}_t | \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}) = \Pr(\mathbf{f}_t | \mathbf{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}) \Pr(\mathbf{C}_t | \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}),$$

and then to make the assumption that the confidence measure produced by the formant analyser is a reliable estimate, hence  $\Pr(\mathbf{C}_t | \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}) = \Pr(\mathbf{C}_t) = 1$ , since the confidence

measure is clearly independent of the output distribution parameters.

It can now be argued that the model proposed so far is mathematically the same as a noisy channel model, and that in practice it is easier to think of it in these terms. The state output distribution emits a value,  $\mathbf{y}_t$ , which then passes through a noisy channel with zero mean and covariance  $\mathbf{C}_t$ ;  $\mathbf{f}_t$  is then the noisy observation. The expression of interest is clearly

$$\Pr(\mathbf{f}_t | \mathbf{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}),$$

which is the same as before, but without the prior on  $\mathbf{C}_t$ .

To evaluate this expression, we must acknowledge that the measured vector,  $\mathbf{f}_t$ , depends upon the unknown output vector  $\mathbf{y}_t$ , and this vector must be integrated out:

$$\Pr(\mathbf{f}_t | \mathbf{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}) = \int_{\mathbb{R}^n} d\mathbf{y}_t \Pr(\mathbf{f}_t | \mathbf{y}_t, \mathbf{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}) \Pr(\mathbf{y}_t | \mathbf{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}),$$

where  $\mathbb{R}^n$  denotes the  $n$ -dimensional Euclidean space of possible observations. Observing some obvious independencies and substituting normal distributions,

$$\Pr(\mathbf{f}_t | \mathbf{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}) = \int_{\mathbb{R}^n} d\mathbf{y}_t \mathcal{N}(\mathbf{f}_t; \mathbf{y}_t, \mathbf{C}_t) \mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}).$$

This form is very intuitive, it just states that the output probability should be evaluated for all possible values of the feature vector, weighted by the formant analyser's belief of each value. Given that  $\mathcal{N}(\mathbf{f}_t; \mathbf{y}_t, \mathbf{C}_t) \equiv \mathcal{N}(\mathbf{y}_t; \mathbf{f}_t, \mathbf{C}_t)$ , the integral is the convolution of two normal distributions. It can be shown that the variances simply add, the result being

$$\Pr(\mathbf{f}_t | \mathbf{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}) = \mathcal{N}(\mathbf{f}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t} + \mathbf{C}_t).$$

So, to incorporate the formant variances in recognition, we simply add the appropriate confidence variance to that of the output distribution. This result is intuitively pleasing: For high confidence (low variance), the usual expression applies, and for low confidence the output distribution widens to equally favour all output values.

### 3.2. Re-estimation

The re-estimation problem is to find a set of parameters  $\boldsymbol{\lambda}$  which maximises the likelihood  $\Pr(\mathbf{O} | \boldsymbol{\lambda})$  of an observation sequence  $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_T$ .  $\boldsymbol{\lambda}$  consists of an  $S \times S$  transition probability matrix  $\mathbf{A}$ , and means and covariance matrices  $\boldsymbol{\mu}_i$  and  $\boldsymbol{\Sigma}_i$ , where  $i = 1, \dots, S$ . Substituting the pair  $\{\mathbf{f}_t, \mathbf{C}_t\}$  for  $\mathbf{o}_t$ , the probability of the observation is

$$\Pr(\mathbf{O} | \boldsymbol{\lambda}) = \sum_s a_{s_0} \prod_{t=1}^T a_{s_{t-1}s_t} \mathcal{N}(\mathbf{f}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t} + \mathbf{C}_t).$$

Following Liporace's interpretation of Baum's method [2], we define an auxiliary function  $Q(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}})$ :

$$Q(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}) = \sum_s \Pr(\mathbf{O}, \mathbf{s} | \boldsymbol{\lambda}) \log \Pr(\mathbf{O}, \mathbf{s} | \bar{\boldsymbol{\lambda}}),$$

which has the property that

$$Q(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}) > Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}) \Rightarrow \Pr(\mathbf{O} | \bar{\boldsymbol{\lambda}}) > \Pr(\mathbf{O} | \boldsymbol{\lambda}).$$

Expanding the  $\bar{\boldsymbol{\lambda}}$  portion of  $Q$  and rearranging the final term to isolate the parameters to be re-estimated,

$$\begin{aligned} Q(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}) = & \sum_s \Pr(\mathbf{O}, \mathbf{s} | \boldsymbol{\lambda}) \times \left[ \log \bar{a}_{s_0} \right. \\ & + \sum_{t=1}^T \left\{ \log \bar{a}_{s_{t-1}s_t} - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\bar{\boldsymbol{\Sigma}}_{s_t} + \mathbf{C}_t| \right. \\ & \left. \left. - \frac{1}{2} (\mathbf{f}_t - \bar{\boldsymbol{\mu}}_{s_t})' (\bar{\boldsymbol{\Sigma}}_{s_t} + \mathbf{C}_t)^{-1} (\mathbf{f}_t - \bar{\boldsymbol{\mu}}_{s_t}) \right\} \right]. \end{aligned}$$

It is clear that the re-estimation equations for the transition probabilities will be unchanged from the standard ones. The means and covariances, however, are likely to be different. First the means: Given that

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{y} - \mathbf{x})' \mathbf{A} (\mathbf{y} - \mathbf{x}) = -\mathbf{A} (\mathbf{y} - \mathbf{x}) - \mathbf{A}' (\mathbf{y} - \mathbf{x})$$

for any general matrix  $\mathbf{A}$ , and the covariance term is symmetric,

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}})}{\partial \bar{\boldsymbol{\mu}}_j} = & - \sum_s \Pr(\mathbf{O}, \mathbf{s} | \boldsymbol{\lambda}) \sum_{\{t: s_t=j\}} (\bar{\boldsymbol{\Sigma}}_j + \mathbf{C}_t)^{-1} (\mathbf{f}_t - \bar{\boldsymbol{\mu}}_j), \end{aligned}$$

Interchanging the order of summation and equating to zero,

$$\sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathbf{O}, \mathbf{s} | \boldsymbol{\lambda}) (\bar{\boldsymbol{\Sigma}}_j + \mathbf{C}_t)^{-1} (\mathbf{f}_t - \bar{\boldsymbol{\mu}}_j) = 0,$$

Following Liporace, we would be able to pre-multiply by the inverse of the matrix term. Here, however,  $\mathbf{C}_t$  is frame dependent and must remain. Rearranging yields the re-estimation formula for the mean:

$$\begin{aligned} \bar{\boldsymbol{\mu}}_j = & \left( \sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathbf{O}, \mathbf{s} | \boldsymbol{\lambda}) (\bar{\boldsymbol{\Sigma}}_j + \mathbf{C}_t)^{-1} \right)^{-1} \\ & \times \sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathbf{O}, \mathbf{s} | \boldsymbol{\lambda}) (\bar{\boldsymbol{\Sigma}}_j + \mathbf{C}_t)^{-1} \mathbf{f}_t. \end{aligned}$$

We assume that the current value of the covariance matrix can be used, instead of the re-estimate. We also note that, in the fully multivariate case, this expression requires a matrix inversion for each frame.

Now consider the covariance re-estimation. Liporace differentiates with respect to the inverse, but here it is more

convenient to use the matrix itself:

$$\begin{aligned} \frac{\partial Q(\lambda, \bar{\lambda})}{\partial \bar{\Sigma}_j} = & \\ & - \frac{1}{2} \sum_s \Pr(\mathcal{O}, \mathbf{s} | \lambda) \sum_{\{t: s_t=j\}} \left\{ \frac{\partial}{\partial \bar{\Sigma}_j} \log |\bar{\Sigma}_j + \mathbf{C}_t| \right. \\ & \left. + \frac{\partial}{\partial \bar{\Sigma}_j} (\mathbf{f}_t - \bar{\boldsymbol{\mu}}_j)' (\bar{\Sigma}_j + \mathbf{C}_t)^{-1} (\mathbf{f}_t - \bar{\boldsymbol{\mu}}_j) \right\}. \end{aligned}$$

Taking each term separately,

$$\frac{\partial}{\partial \bar{\Sigma}_j} \log |\bar{\Sigma}_j + \mathbf{C}_t| = (\bar{\Sigma}_j + \mathbf{C}_t)^{-1}.$$

Strictly, when differentiating with respect to a symmetric matrix, the off diagonal elements of the result should be doubled [4]. In this case, however, this result is to be combined with another where the same effect happens, and it is both consistent and more readable to ‘ignore’ this effect. Liporace’s derivation omits this caveat, though his results are valid for the same reason. It can be shown with reference to [3] that

$$\frac{\partial}{\partial \mathbf{A}} \mathbf{x}'(\mathbf{A} + \mathbf{B})^{-1} \mathbf{x} = -[(\mathbf{A} + \mathbf{B})^{-1}]' \mathbf{x} \mathbf{x}' [(\mathbf{A} + \mathbf{B})^{-1}]'.$$

So, denoting  $\mathbf{f}_t - \bar{\boldsymbol{\mu}}_j$  by  $\mathbf{x}$ , interchanging the order of summation as before and equating to zero yields

$$\begin{aligned} \sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathcal{O}, \mathbf{s} | \lambda) \left[ (\bar{\Sigma}_j + \mathbf{C}_t)^{-1} \right. \\ \left. - (\bar{\Sigma}_j + \mathbf{C}_t)^{-1} \mathbf{x} \mathbf{x}' (\bar{\Sigma}_j + \mathbf{C}_t)^{-1} \right] = 0. \quad (1) \end{aligned}$$

At this stage, it is clear that  $\bar{\Sigma}_j$  cannot be isolated, and it is necessary to make an approximation. Two alternative approximations are proposed, as described below.

### 3.2.1. Method 1

Equation 1 can also be written,

$$\begin{aligned} \sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathcal{O}, \mathbf{s} | \lambda) \left[ \right. \\ \left. (\bar{\Sigma}_j + \mathbf{C}_t)^{-1} (\bar{\Sigma}_j + \mathbf{C}_t - \mathbf{x} \mathbf{x}') (\bar{\Sigma}_j + \mathbf{C}_t)^{-1} \right] = 0. \end{aligned}$$

We now assume that  $\mathbf{C}_t$  is independent of time for a given state, that is, it can be assumed constant for the duration of the state. This approximation is not unreasonable because the confidence with which the formant frequencies are estimated will generally be similar for all the feature vectors corresponding to any one model state. The two inverse terms can now be brought outside the summation, and the expression can then be pre- and post-multiplied by the inverse of those terms leaving a single instantiation of  $\bar{\Sigma}_j$ :

$$\sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathcal{O}, \mathbf{s} | \lambda) (\bar{\Sigma}_j + \mathbf{C}_t - \mathbf{x} \mathbf{x}') = 0.$$

Rearranging,

$$\bar{\Sigma}_j = \frac{\sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathcal{O}, \mathbf{s} | \lambda) [\mathbf{x} \mathbf{x}' - \mathbf{C}_t]}{\sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathcal{O}, \mathbf{s} | \lambda)}.$$

This approximation appears to have a problem: Where  $\mathbf{C}_t$  is large, the term in square brackets will not be positive definite, which is one of the conditions cited by Liporace for the re-estimation to be valid. A remedy is to simply ignore the contribution of frames for which this term is not positive definite, that is, the sum of the eigenvalues is not positive. The effect of this is that the system is not trained on low confidence frames, which is entirely reasonable. For states where one or more frame elements are always low confidence, we suggest that this will be true in recognition too, and hence the  $\mathbf{C}_t$  term will dominate there also. In the particular case where the covariance is assumed diagonal, the individual elements of the term in square brackets can be handled individually.

### 3.2.2. Method 2

Starting again from equation 1, notice that the first term in the squares brackets can be written

$$(\bar{\Sigma}_j + \mathbf{C}_t)^{-1} = \bar{\Sigma}_j^{-1} (\mathbf{I} + \mathbf{C}_t \bar{\Sigma}_j^{-1})^{-1}. \quad (2)$$

Hence, by substituting equation 2 into equation 1, and pre- and post-multiplying both sides by  $\bar{\Sigma}_j$ , a term in  $\bar{\Sigma}_j$  can be isolated. This means that the equation can be rearranged thus:

$$\begin{aligned} \bar{\Sigma}_j = & \\ & \left( \sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathcal{O}, \mathbf{s} | \lambda) (\mathbf{I} + \mathbf{C}_t \bar{\Sigma}_j^{-1})^{-1} \bar{\Sigma}_j \right)^{-1} \quad (3) \\ & \times \sum_{t=1}^T \sum_{\{s: s_t=j\}} \Pr(\mathcal{O}, \mathbf{s} | \lambda) \\ & \bar{\Sigma}_j (\bar{\Sigma}_j + \mathbf{C}_t)^{-1} \mathbf{x} \mathbf{x}' (\bar{\Sigma}_j + \mathbf{C}_t)^{-1} \bar{\Sigma}_j. \end{aligned}$$

If it is assumed that  $\bar{\Sigma}_j$  terms on the right hand side can be replaced by their previous values, then equation 3 constitutes a re-estimation equation for  $\bar{\Sigma}_j$ .

## 4. COROLLARY

The problem as described is applicable to any feature set which is subject to additive, time varying Gaussian noise. A particular special case is that where the uncertainty (noise) can be assumed constant with time. Practically, this means that  $\mathbf{C}$  is no longer dependent upon  $t$ , and certain matrix terms in the re-estimation equations become independent of the summation and cancel. In particular, the first re-estimate of the covariance above ceases to be an approximation, and the re-estimate of the mean reverts to the same as that for the conventional noiseless case.

## 5. EXPERIMENTS

### 5.1. Method

The new method for incorporating formant confidence measures in both training and recognition was tested using the same speaker-independent connected-digit recognition task with three-state phone models as was used in earlier studies [1]. As with the previous experiments, the baseline feature set comprised the first eight mel-cepstrum coefficients and an overall energy feature. The performance of this feature set was compared with one in which coefficients 6, 7 and 8 were replaced by three formant features for describing fine spectral detail. In the case of the formant features, the confidence measures were incorporated first in recognition and then also in training, testing both of the approximations suggested in the previous section for the re-estimation of the model variances. For both training algorithms, it was verified experimentally from the training-set probabilities that the re-estimation process converged after a few iterations. For all model sets, a total of ten iterations were performed before testing the models in recognition.

Alternative formant sets arising from labelling ambiguity were optionally accommodated in training and recognition, simply by choosing the formant set which gave the highest HMM emission probability for each frame and model state. Results using the confidences and alternative formant sets were compared with those obtained when no special treatment was given to the formant features.

### 5.2. Results and Discussion

From the results shown in Table 1 it can be seen that the formant features gave very poor performance unless the degree of confidence in their measurement accuracy was taken into account. When the formant features were not given special treatment, there were serious problems with insertion errors. These errors were caused by mismatches between the formant frequencies in the non-speech models with those measured for the non-speech regions of the test data. These errors disappeared when the confidence measure was incorporated in recognition.

A small additional benefit was obtained by also incorporating the confidence measure in training, with very similar results being obtained for the two suggested approaches to training the model variances. In all cases, further small improvements in recognition performance were obtained by including alternative formant sets. The lowest error-rate of

2.5% that was achieved with the formants demonstrates a substantial improvement over the figure of 4.0% that was obtained when using only mel-cepstrum features, for the same total number of features.

These digit-recognition experiments have provided a good basis for initial comparisons, and experiments are now in progress to evaluate performance on the more demanding task of phone recognition using the TIMIT database.

## 6. CONCLUSIONS

We have shown that formant frequency estimates with confidence levels can be interpreted probabilistically, and that this interpretation leads to theoretically justifiable variants of the standard HMM recognition and re-estimation equations. Further, the theoretical results have been evaluated experimentally and shown to work in practice. Considerable recognition advantages have been demonstrated from incorporating formant features in this way, in comparison with using only cepstrum features.

It is planned to incorporate the formant representation into a segmental modelling paradigm to model formant trajectories, and then to progress towards developing an appropriate underlying model of time evolving speech characteristics.

## 7. REFERENCES

- [1] John N. Holmes, Wendy J. Holmes, and Philip N. Garner. Using formant frequencies in speech recognition. In *Proceedings EUROSpeech'97*, volume 4, pages 2083–2086, September 1997.
- [2] Louis A. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory*, IT-28(5):729–734, September 1982.
- [3] Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. Wiley series in probability and mathematical statistics. John Wiley and sons, 1988.
- [4] Shayle R. Searle. *Matrix algebra useful for statistics*. Wiley series in probability and mathematical statistics. John Wiley and sons, 1982.

©British Crown Copyright 1997/DERA  
Published with the permission of the Controller of Her  
Britannic Majesty's Stationery Office.

Model Set	%Subs.	%Del.	%Ins.	%Err.
8 cepstrum features+energy	2.8	1.0	0.2	4.0
5 cepstrum features+energy+3 formants	5.2	1.0	10.2	16.4
Add formant confidence measure (recognition only)	2.1	0.7	0.2	3.0
Also include second choice formants in recognition	2.1	0.4	0.4	2.9
Add confidence measure in training (method 1)	2.0	0.6	0.2	2.8
Also include second choice formants (training and recognition)	1.9	0.6	0.1	2.6
Add confidence measure in training (method 2)	2.0	0.6	0.2	2.8
Also include second choice formants (training and recognition)	1.8	0.6	0.1	2.5

Table 1: Connected-digit recognition performance for different feature sets.

# A DIFFERENTIAL SPECTRAL VOICE ACTIVITY DETECTOR

*Philip N. Garner, Toshiaki Fukada and Yasuhiro Komori*

Canon Inc.

3-30-2 Shimomaruko, Ota-ku, Tokyo 146-8501, Japan.

Email: philip.garner@canon.co.jp, fukada.toshiaki@canon.co.jp, komori.yasuhiro@canon.co.jp

## ABSTRACT

The Voice Activity Detection (VAD) problem is placed into a decision theoretic framework, and the Gaussian VAD model of Sohn *et al.* is then shown to fit well with the framework. It is argued that the Gaussian model can be made more robust to correlation and expected spectral shapes of speech and noise by using a differential spectral representation. Such a model is formulated theoretically. The differential spectral VAD is then shown by experiment to be consistently superior to the basic Gaussian VAD in a speech recognition setting, especially for noisy environments.

## 1. INTRODUCTION

Voice Activity Detection (VAD) is important in various applications involving speech. Perhaps the most common application is in telecommunications, where the main reason for the VAD is to save bandwidth by not transmitting non-speech portions of the input signal. Marzinzik and Kollmeier [1] present a useful recent review of the subject.

We are interested in (VAD) for the purpose of Automatic Speech Recognition (ASR) in general, and noise robust ASR in particular. VAD is important in ASR because it distinguishes the non-speech portions at the beginning and end of an utterance from the utterance itself. In doing this, the VAD ensures that the decoder, which is computationally intensive, only runs when necessary. This point is particularly important in embedded applications, where processing power is limited. The main difference between a VAD used in telecommunications and one used in ASR is that the latter typically uses a state machine in order to avoid false detections and to remain active during speech pauses.

A VAD working in the spectral domain, and with an appealing statistical basis, has been introduced by Sohn *et al.* [2, 3]. This spectral VAD has been shown to be superior to three standard VADs (QCELP, EVRC and G.729B) in a telecommunications environment. That result is reinforced in a comparison by Stadermann *et al.* [4], who find the spectral VAD to be superior to baselines based on frame energy and spectral entropy. The work has also been extended by Cho *et al.* [5], who show that smoothing can alleviate problems with errors in the end of speech region.

The spectral VAD of Sohn *et al.* is based on a simple Gaussian assumption. This basic Gaussian model has two distinct problems:

1. The model has no knowledge of the spectral shape of either the speech or the noise. It is well known, however, that speech has distinct spectral peaks (formants). Conversely, many types of noise have a smooth spectral shape.

2. From a purely statistical point of view, the model assumes that adjacent spectral bins are uncorrelated. This is not true, especially for speech, and even more especially for observations from the overlapped triangular mel-spaced filterbank typical of current ASR systems.

In this paper, we evaluate the basic Gaussian VAD algorithm described above in an ASR context. We then argue that a differential spectral representation can minimize the effects of the two problems described above. We derive a differential spectral version of the Gaussian VAD, and show that it leads to improved performance.

## 2. BACKGROUND THEORY

### 2.1. Decision theoretic framework

Define a boolean variable or hypothesis  $\mathcal{H}$ , which can take values 0 and 1.  $\mathcal{H} = 0$  indicates non-speech and  $\mathcal{H} = 1$  indicates the presence of speech. A VAD produces an estimate (or choice),  $\hat{\mathcal{H}}$ , given some observation. For this derivation, assume that the observation is the (complex) spectrum  $\mathfrak{s}$ .

The above leads to a simple decision theoretic formulation: Define a loss or cost function,  $C(\mathcal{H}, \hat{\mathcal{H}})$ , that attaches a cost to each combination of  $\mathcal{H}$  and  $\hat{\mathcal{H}}$ . Typically, the cost should be low for a correct classification, and high for an incorrect one. The expected costs of the two possible classifications are then

$$E(C(\mathcal{H}, 0) | \mathfrak{s}) = \sum_{\mathcal{H}} C(\mathcal{H}, 0) P(\mathcal{H} | \mathfrak{s}), \quad (1)$$

$$E(C(\mathcal{H}, 1) | \mathfrak{s}) = \sum_{\mathcal{H}} C(\mathcal{H}, 1) P(\mathcal{H} | \mathfrak{s}). \quad (2)$$

We can now choose the classification,  $\hat{\mathcal{H}}$ , that has the smaller expected cost; that is: Choose  $\hat{\mathcal{H}} = 1$  if

$$\sum_{\mathcal{H}} C(\mathcal{H}, 1) P(\mathcal{H} | \mathfrak{s}) < \sum_{\mathcal{H}} C(\mathcal{H}, 0) P(\mathcal{H} | \mathfrak{s}). \quad (3)$$

Expanding the summations and rearranging,

$$\frac{P(\mathcal{H} = 1 | \mathfrak{s})}{P(\mathcal{H} = 0 | \mathfrak{s})} > \frac{C(0, 1) - C(0, 0)}{C(1, 0) - C(1, 1)}. \quad (4)$$

Given that we will assume a model for the generation of  $\mathfrak{s}$ , it is useful to apply Bayes's theorem to the conditional probabilities in equation 4. Notice that the evidence (denominator of Bayes's

theorem) term cancels, giving

$$\underbrace{\frac{p(\mathbf{s} | \mathcal{H} = 1)}{p(\mathbf{s} | \mathcal{H} = 0)}}_{\text{Likelihood ratio, } L(\mathbf{s})} \cdot \underbrace{\frac{P(\mathcal{H} = 1)}{P(\mathcal{H} = 0)}}_{\text{Prior ratio}} > \underbrace{\frac{C(0, 1) - C(0, 0)}{C(1, 0) - C(1, 1)}}_{\text{Cost ratio}}, \quad (5)$$

where we refer to the terms as indicated.

The prior ratio and cost ratio can be set to unity given the following broad assumptions:

- The likelihood of an observation  $\mathbf{s}$  being speech is as likely as it being non-speech.
- The cost of an accurate classification is zero, and the costs of the two inaccurate classifications are identical.

Of course, the above assumptions may not be true for a given scenario, in which case the terms can be set accordingly. The combination of cost ratio and prior ratio into a single threshold term yields the likelihood ratio test used by Sohn *et al.*. The advantage of the decision theoretic approach is that it gives some insight into what the threshold should be.

## 2.2. Gaussian model

Broadly following Sohn *et al.* [3], but with a minor change of notation to allow subscripts to refer to vector elements, assume that both the speech and noise can be modeled by Gaussian distributions (more accurately, the real and imaginary components of each spectral bin are i.i.d. Gaussian). This is identical to the assumption made in the Ephraim Malah formulation for speech enhancement [6]. We define two probability distributions:

$$p(\mathbf{s} | \mathcal{H} = 0) = \prod_{k=1}^S \frac{1}{\pi \mu_k} \exp\left(-\frac{s_k^2}{\mu_k}\right), \quad (6)$$

$$p(\mathbf{s} | \mathcal{H} = 1) = \prod_{k=1}^S \frac{1}{\pi(\lambda_k + \mu_k)} \exp\left(-\frac{s_k^2}{\lambda_k + \mu_k}\right), \quad (7)$$

where  $\mathbf{s}$  is the  $S$  dimensional complex spectrum observation,  $s_k$  is the magnitude of the  $k^{\text{th}}$  element of  $\mathbf{s}$ ,  $\lambda_k$  is the variance of the  $k^{\text{th}}$  dimension of the speech signal and  $\mu_k$  is the variance of the  $k^{\text{th}}$  dimension of the noise signal. All of the above is for a single frame, although the  $f$  subscript is omitted for clarity. Equation 7 follows from that fact that the sum of two Gaussian random variates is Gaussian with variance equal to the sum of the individual variances.

Substituting equations 6 and 7 into equation 5 gives a VAD likelihood ratio of

$$L(\mathbf{s}) = \prod_{k=1}^S \frac{\mu_k}{\lambda_k + \mu_k} \exp\left(\frac{\lambda_k}{\lambda_k + \mu_k} \cdot \frac{s_k^2}{\mu_k}\right). \quad (8)$$

Notice that equation 8 is defined in terms of spectral power measures, even though the assumptions so far are based on complex spectrum.

## 2.3. Correction for correlation

When taking a product of probabilities known to be correlated, it is normal to make a simple correction for the correlation in the form of a weighted geometric mean,

$$p(\mathbf{s}) = \prod_{k=1}^S p(s_k)^{\frac{1}{\kappa^S}}, \quad (9)$$

where  $\kappa$  is an optimised constant analogous to the language model match factor in ASR. Sohn *et al.* do this implicitly by taking the unweighted geometric mean ( $\kappa = 1$ ), although in this framework that is an extreme solution and represents absolute correlation between bins.  $\kappa = 1/S$  represents complete independence.

## 3. DIFFERENTIAL SPECTRAL VAD

We suggest that the single zero high-pass filter (HPF),

$$s_k^{2'} = s_{k+1}^2 - s_k^2 \quad 1 \leq k < S, \quad (10)$$

applied in the frequency dimension of each power spectral frame will tackle the problems highlighted in section 1 as follows:

1. The HPF will map the smooth spectrum associated with noise, especially the flat spectrum of white noise or impulse noise, to a flatter spectrum centered around zero. This is much closer to the spectrum of silence.
2. The subtraction will reduce or eliminate the correlation between adjacent spectral bins.

In fact, the decorrelation effect has been demonstrated in the context of robust ASR by Nadeu *et al.* [7], who show that such a filter can be used in place of the cosine transform normally used in ASR.

In the VAD context, however, we require a probability distribution associated with the filter. This is derived as follows:

First, notice that the single zero filter of equation 10 corresponds to a probabilistic change of variable with  $1 \leq k < S$  and an integral over  $s_S^2$ . This integral turns out to be highly non-trivial. Instead, we decimate the above substitution as follows, allowing the problem to be solved as  $S/2$  identical and much simpler integrals:

$$s_k^{2'} = s_{2k}^2 - s_{2k-1}^2 \quad 1 \leq k \leq S/2. \quad (11)$$

In this case, the length of the resulting feature vector is  $S/2$  instead of  $S - 1$ . For the rest of the derivation in this section, as the integrals are identical, we simply consider the case where  $k = 1$ .

Second, given that the distribution of the complex spectrum is Gaussian, it can be shown by change of variable that the distribution of spectral power is the exponential distribution,

$$p(s^2 | v) = \frac{1}{v} \exp\left(-\frac{s^2}{v}\right), \quad (12)$$

where  $v$  is a variance parameter to be substituted later. It follows that the joint distribution of two exponentially distributed observations is

$$p(s_1^2, s_2^2 | v_1, v_2) = \frac{1}{v_1 v_2} \exp\left(-\frac{s_1^2}{v_1} - \frac{s_2^2}{v_2}\right). \quad (13)$$

The PDF of the filtered signal arises from changing one of the variables to  $z = s_2^2 - s_1^2$  and integrating out the other variable. To perform the integral, notice that in the case where  $z \geq 0$ ,  $s_2^2 \geq z$  and  $s_1^2 \geq 0$ . Also, when  $z \leq 0$ ,  $s_1^2 \geq -z$  and  $s_2^2 \geq 0$ . This suggests the use of two different integrals:



In the case where  $z \geq 0$ ,

$$\begin{aligned}
p(z | v_1, v_2) &= \int_0^\infty ds_1^2 p(s_1^2) p(z + s_1^2) \\
&= \int_0^\infty ds_1^2 \frac{1}{v_1 v_2} \exp\left(-\frac{s_1^2}{v_1} - \frac{z + s_1^2}{v_2}\right) \\
&= \frac{1}{v_1 v_2} \exp\left(-\frac{z}{v_2}\right) \\
&\quad \times \int_0^\infty ds_1^2 \exp\left(-s_1^2 \left[\frac{1}{v_1} + \frac{1}{v_2}\right]\right) \\
&= \frac{1}{v_1 v_2} \exp\left(-\frac{z}{v_2}\right) \cdot \frac{v_1 v_2}{v_1 + v_2}.
\end{aligned} \tag{14}$$

Similarly, in the case where  $z \leq 0$ ,

$$\begin{aligned}
p(z | v_1, v_2) &= \int_0^\infty ds_2^2 p(s_2^2) p(s_2^2 - z) \\
&= \frac{1}{v_1 v_2} \exp\left(\frac{z}{v_1}\right) \cdot \frac{v_1 v_2}{v_1 + v_2}.
\end{aligned} \tag{15}$$

Substituting back for  $z$ , and combining the two results,

$$\begin{aligned}
p(s_2^2 - s_1^2 | v_1, v_2) &= \\
&\begin{cases} \frac{1}{v_1 + v_2} \exp\left(-\frac{s_2^2 - s_1^2}{v_2}\right) & \text{if } s_2^2 \geq s_1^2, \\ \frac{1}{v_1 + v_2} \exp\left(-\frac{s_1^2 - s_2^2}{v_1}\right) & \text{if } s_2^2 \leq s_1^2. \end{cases}
\end{aligned} \tag{16}$$

Note that both expressions are identical when  $s_1^2 = s_2^2$ .

The likelihood ratio follows easily from equation 16 by substituting for  $v_1$  and  $v_2$ . Assuming for simplicity that  $s_2^2 > s_1^2$ , the likelihood ratio is the ratio of the following two equations:

$$\begin{aligned}
p(s_2^2 - s_1^2 | \mathcal{H} = 1) &= \\
&\frac{1}{\mu_1 + \lambda_1 + \mu_2 + \lambda_2} \exp\left(-\frac{s_2^2 - s_1^2}{\mu_2 + \lambda_2}\right),
\end{aligned} \tag{17}$$

and,

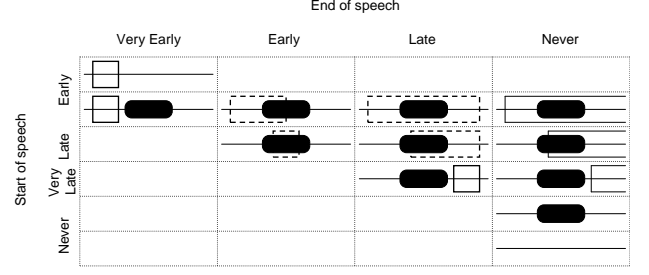
$$p(s_2^2 - s_1^2 | \mathcal{H} = 0) = \frac{1}{\mu_1 + \mu_2} \exp\left(-\frac{s_2^2 - s_1^2}{\mu_2}\right), \tag{18}$$

which evaluates to

$$\begin{aligned}
L(s_2^2 - s_1^2) &= \frac{\mu_1 + \mu_2}{\mu_1 + \lambda_1 + \mu_2 + \lambda_2} \\
&\quad \times \exp\left(-\frac{s_2^2 - s_1^2}{\mu_2 + \lambda_2} + \frac{s_2^2 - s_1^2}{\mu_2}\right), \\
&= \frac{\mu_1 + \mu_2}{\mu_1 + \lambda_1 + \mu_2 + \lambda_2} \\
&\quad \times \exp\left(\frac{s_2^2 - s_1^2}{\mu_2} \cdot \frac{\lambda_2}{\mu_2 + \lambda_2}\right).
\end{aligned} \tag{19}$$

The full likelihood ratio is the product of this expression applied to each pair of spectral bins,

$$L(\mathbf{s}) = \prod_{k=1}^{S/2} L(s_{2k}^2 - s_{2k-1}^2). \tag{20}$$



**Fig. 1.** Classification of VAD start and end times. The dark portion represents speech, the box represents the VAD result.

## 4. EVALUATION

### 4.1. Testing data

The VADs were evaluated using an in-house database, some aspects of which were designed specifically for VAD evaluation. The database consists of 14 speakers (7 male and 7 female) each speaking 40 utterances in each of 6 different environments. This is 3360 utterances in total. The utterances are isolated Japanese city names, but are each 5 seconds in length. Typically, the first 2 seconds are background noise, the utterance itself is one second or less, and the final 2 seconds are background noise. The data have been manually marked up with the speech start and end times. The data were recorded on a portable (PDA-like) device using an ear-mounted microphone, the actual microphone being close to the speaker's cheek.

Five of the six environments were chosen to be representative of those where the portable device might be used:

1. The laboratory sound-proof room.
2. A large open-plan office with carpets and fans.
3. A reverberant but open and quiet company lobby.
4. A cafeteria at lunch-time with constant babble noise.
5. A busy suburban street with occasional traffic.
6. A quieter, more open, outdoor area on a windy day.

The average signal to noise ratio for each environment is shown in table 1.

### 4.2. Evaluation metric

The main evaluation metric consisted of a classification of each utterance into one of the states indicated in figure 1. These are based on a combination of the speech start and end times, and can be thought of as a variation of the classes used by Rosca *et al.* [8]. The four classifications drawn with dashed lines represent the VAD working well, or in such a way that can be corrected using wide margins. The bottom result in the right-most column represents a correct non-detection of an empty utterance, one of which exists (accidentally) in our database. The seven other classifications are certainly errors, being either insertions, deletions or the offset time not being detected in the recording.

The right-most column of figure 1 provides a useful metric for optimizing  $\kappa$ : too small a value leads to large likelihoods and

**Table 1.** Error rate (%) for four VAD configurations. Also shown is the SNR for each environment, and the optimized value of  $\kappa$  for each VAD configuration.

	SNR (dB)	Power		Mel	
		Gauss	Diff.	Gauss	Diff.
1 (clean)	28.5	0.4	0.4	0.2	0.2
2 (office)	24.7	1.8	1.8	1.3	1.0
3 (lobby)	24.1	0.7	0.5	0.4	0.2
4 (cafe)	16.6	9.8	9.6	4.6	3.8
5 (street)	15.8	6.3	4.1	3.6	3.4
6 (outside)	21.4	6.3	5.2	8.9	5.5
$\kappa$		2.5	3.0	1.0	1.0

missing end times. Too large a value, however, leads to deletions. For ASR, we favour insertions over deletions as insertions can be handled using garbage modeling. Missing end times, however, are particularly bad as they cause the recogniser to “hang” and ultimately give an errorful recognition.

### 4.3. VAD construction

The VAD is inserted into the spectral part of the normal signal processing chain used in ASR. In this case, the signal is sampled at 11.025 KHz and pre-emphasized. Overlapping frames of 256 samples are then taken every 10 ms to form a 128 bin power spectrum (the bin at  $\pi$  is discarded). The power spectrum is transformed into 32 mel spaced bins using half overlapping triangular filters.

The noise variance from the previous frame,  $\mu_{f-1}$ , is used in the likelihood calculation, and is then updated using a slightly modified version of the estimator described by Sohn *et al.* [2],

$$\hat{\mu}_f = \frac{1 - \rho_\mu}{1 + L(\mathfrak{s})} \mathfrak{s}_f^2 + \frac{\rho_\mu + L(\mathfrak{s})}{1 + L(\mathfrak{s})} \hat{\mu}_{f-1}, \quad (21)$$

where  $\rho_\mu = 0.95$ , and  $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_S)^T$ . The speech variances,  $\lambda_k$ , are estimated using power spectral subtraction as suggested in [2], except with the usual over-subtraction and flooring. The VAD was found not to be sensitive to the over-subtraction and flooring values, but note that the flooring means that equation 8 does not reach it’s minimum value of 1. For this reason, the cost ratio of equation 5 was set a little above 1 (actually 2.5).

The actual start and end points of the speech were determined using a simple state machine that requires at least 10 frames indicated to be speech in a 1 second window in order to transition to the speech state, and 40 frames of contiguous non-speech to transition into the non-speech state.

### 4.4. Results

The original Gaussian based VAD, and the differential spectral VAD were tested in both power spectral and mel spectral domains. In each case, the parameter  $\kappa$  was adjusted manually to minimize the number of deletions and missing end times as described in section 4.2. The results are shown in table 1.

The first 3 environments are relatively noise-free; there is no significant difference resulting from the choice of VAD. There is, however, a slight bias in favour of using the mel domain. The latter 3 environments are comparatively noisy, and show more variation. In particular, the mel domain is more robust to the babble noise of

the cafeteria. The power spectral domain, however, is more suited to the outdoor wind noise.

Broadly, the differential VAD produces fewer errors than the equivalent non-differential formulation. This confirms the utility of the differential spectral approach. We have also confirmed that this improved VAD performance leads directly to improved speech recognition performance on the same database.

Finally, one obvious difference between the Gaussian and differential VADs is that the former uses a probability for each spectral bin, whereas the latter uses a probability for each pair of bins. In order to confirm that the advantage of the differential VAD is not simply through using half the parameters, we constructed a comparable Gaussian VAD by averaging adjacent bins. This approach only contributed detrimentally to the performance.

## 5. CONCLUSION

We have placed a spectral VAD into a rigorous decision theoretic framework, and evaluated it in an ASR environment. In order to optimize it to an ASR feature space, and make it robust to noises with smooth spectra, we have re-formulated it as a differential spectral VAD, again in a rigorous statistical manner. We have shown the differential spectral formulation to be superior to the basic Gaussian for an ASR application.

## 6. REFERENCES

- [1] M. Marzinzik and B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, February 2002.
- [2] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1998, pp. 365–368.
- [3] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, January 1999.
- [4] J. Stadermann, V. Stahl, and G. Rose, “Voice activity detection in noisy environments,” in *Proceedings of EUROSPEECH*, Scandinavia, 2001.
- [5] Y. D. Cho, K. Al-Naimi, and A. Kondoz, “Improved voice activity detection based on a smoothed statistical likelihood ratio,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2001.
- [6] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.
- [7] C. Nadeu, D. Macho, and J. Hernando, “Time & frequency filtering of filter bank energies for robust HMM speech recognition,” *Speech Communication*, vol. 34, pp. 93–114, April 2001.
- [8] J. Rosca, R. Balan, N. P. Fan, C. Beaugeant, and V. Gilg, “Multichannel voice detection in adverse environments,” in *Proceedings of EUSIPCO*, Toulouse, France, September 2002.

# SNR Features for Automatic Speech Recognition

Philip N. Garner

*Idiap Research Institute  
Martigny, Switzerland  
pgarner@idiap.ch*

**Abstract**—When combined with cepstral normalisation techniques, the features normally used in Automatic Speech Recognition are based on Signal to Noise Ratio (SNR). We show that calculating SNR from the outset, rather than relying on cepstral normalisation to produce it, gives features with a number of practical and mathematical advantages over power-spectral based ones. In a detailed analysis, we derive Maximum Likelihood and Maximum a-Posteriori estimates for SNR based features, and show that they can outperform more conventional ones, especially when subsequently combined with cepstral variance normalisation. We further show anecdotal evidence that SNR based features lend themselves well to noise estimates based on low-energy envelope tracking.

## I. INTRODUCTION

An important problem encountered in speech signal processing is that of how to normalise a signal for the effects of noise. In speech enhancement the task is to remove noise from a signal to reproduce the uncorrupted signal such that it is perceived by a listener to be less noisy. In Automatic Speech Recognition (ASR), the task is to reduce the effect of noise on recognition accuracy. In this paper, we will concentrate on the latter (ASR) problem.

Two categories of noise are generally considered: Additive noise is that which represents a distinct signal other than the one of interest. Convolutional noise is that which alters the spectral shape, and can be associated with either the signal of interest, or both the signal and the additive noise.

Cepstral Mean Normalisation (CMN) is a well established technique that compensates for convolutional noise. It is based on the persuasive observation that a linear channel distortion becomes a constant offset in the cepstral domain. CMN also affords some robustness to additive noise. Cepstral Variance Normalisation (CVN) has been observed to provide further noise robustness [1], and the combination of CMN and CVN is now quite ubiquitous in ASR.

Orthogonal to the cepstral normalisation approach, many common practical solutions for additive noise compensation are based on the assumption of a simple additive Gaussian model for both speech and noise in the spectral domain. In ASR, the spectral subtraction approach of Boll [2] is well established, and often used as a means to derive a Wiener filter. In speech enhancement, much work is based on the technique of Ephraim and Malah [3]. Both these techniques have influenced the design of the ETSI standard ASR front-end [4].

Techniques that rely on noise subtraction are dependent upon some means of measuring the background noise in a

signal. Often, it is sufficient to simply average the first few frames of an utterance, however this is not robust to changing noise levels. Ris and Dupont [5] present a survey of methods to measure noise, favouring the low-energy envelope tracking approach of Martin [6]. Lathoud *et al.* [7] present a statistical spectral model that yields both noise and speech estimates.

Cepstral and spectral techniques are often combined. This is a natural approach as, theoretically, the two approaches are designed to tackle different types of noise. For instance, histogram normalisation, a logical progression of CMN/CVN to higher order moments, has been successfully combined with spectral compensation techniques by Segura *et al.* [8]. Lathoud *et al.* [7], who describe their technique as “Unsupervised” spectral subtraction (USS), also report good results in combination with cepstral normalisation.

In this paper, we analyse the relationship between spectral and cepstral normalisation. We first present a simplistic analysis, then a more detailed Bayesian analysis, showing that knowledge of the presence of cepstral compensation should influence the chosen approach to spectral compensation. Theoretical results are evaluated leading to a conclusion that SNR based features represent a theoretically rigorous but computationally simple approach to ASR, and could easily be incorporated into more advanced techniques.

## II. SIMPLISTIC APPROACH TO NOISE

### A. Cepstral Mean Normalisation

In a simplistic, but informative, view of an ASR front-end, an acoustic signal is Fourier transformed to give a vector of spectral coefficients  $(s_1, s_2, \dots, s_F)^T$ . After a linear transform implementing a non-linear frequency warp, the cepstrum is calculated. The cepstrum involves a logarithm followed by another linear transform. In the presence of only convolutional noise,  $(c_1, c_2, \dots, c_F)^T$ , which is multiplicative in the frequency domain, the logarithm becomes

$$\log(c_f s_f) = \log(c_f) + \log(s_f), \quad (1)$$

where  $\log(c_f)$  is constant over time, but  $\log(s_f)$  varies. Hence, subtraction of the cepstral mean results in removal of the constant convolutional noise term. When the filter-bank is considered, the above holds if the  $c_f$  are assumed constant within a given filter-bank bin.

In the presence of only additive noise, the noise is assumed to remain additive after the Fourier transform. In this sense,

the logarithm operation becomes

$$\log(s_f + n_f) = \log(n_f) + \log\left(1 + \frac{s_f}{n_f}\right), \quad (2)$$

where  $(n_1, n_2, \dots, n_F)^\top$  is the noise spectrum. The right hand side of (2) is evident from the Taylor series of  $\log(x + y)$ , and emphasises that CMN would remove the constant term  $\log(n_f)$ .

### B. Properties of SNR features

It appears from the above analysis that, if we use CMN, the features that are presented to the ASR decoder are actually (a linear transform of) the logarithm of one plus the signal to noise ratio (SNR). This will happen even if the additive noise is simply the minimal background noise usually associated with clean recordings. It follows that we could try to calculate the SNR from the outset rather than calculate a spectral power measure and rely on CMN to produce the SNR. A-priori, such an approach has at least two appealing properties:

- 1) The flooring of the logarithm happens naturally. SNR values cannot fall below zero, so the argument of the logarithm is naturally floored at unity.
- 2) SNR is inherently independent of gain associated with microphones and pre-amplifiers.

We will show that SNR is also mathematically appealing.

The approach is analogous to that of Lathoud *et al.* [7]. The only difference is that Lathoud *et al.* explicitly floor the SNR using (in our present notation)

$$\max\left(1, \frac{s_f}{n_f}\right). \quad (3)$$

## III. A MORE RIGOROUS ANALYSIS

In contrast to the previous section, which was left deliberately simplistic, we now present a more rigorous derivation of a SNR based feature. We begin by defining a Gaussian model of speech in noise, and proceed by showing that power spectral subtraction can be seen as a particular maximum-likelihood (ML) solution. We then derive ML and MAP estimators for the SNR.

### A. Gaussian model

Let us assume that a DFT operation produces a vector,  $\mathbf{x}$ , with complex components,  $x_1, x_2, \dots, x_F$ , where the real and imaginary parts of each  $x_f$  are i.i.d. normally distributed with zero mean and variance  $\nu_f$ . That is,

$$f(\mathbf{x}_f | \nu_f) = \frac{1}{\pi\nu_f} \exp\left(-\frac{|\mathbf{x}_f|^2}{\nu_f}\right). \quad (4)$$

In the case where we distinguish two coloured noise signals, a background noise,  $\mathbf{n}$ , and a signal of interest,  $\mathbf{s}$ , typically speech, denote the noise variance as  $\nu$  and the speech variance as  $\varsigma$ . In general, the background noise can be observed in isolation and modelled as

$$f(\mathbf{n}_f | \nu_f) = \frac{1}{\pi\nu_f} \exp\left(-\frac{|\mathbf{n}_f|^2}{\nu_f}\right). \quad (5)$$

The speech, however, cannot normally be observed in isolation. It is always added to noise. When both speech and additive noise are present the variances add, meaning that the total signal,  $\mathbf{t}_f = \mathbf{s}_f + \mathbf{n}_f$ , can be modelled as

$$f(\mathbf{t}_f | \varsigma_f, \nu_f) = \frac{1}{\pi(\varsigma_f + \nu_f)} \exp\left(-\frac{|\mathbf{t}_f|^2}{\varsigma_f + \nu_f}\right). \quad (6)$$

The above model is the basis of the Wiener filter and of the widely used Ephraim-Malah speech enhancement technique [3]. The goal is usually formulated as requiring an estimate of  $\varsigma_f$ ; this proceeds via estimation of  $\nu_f$ .

We assume that an estimate,  $\hat{\nu}$ , of  $\nu$  is available via solution of (5) during, for instance, non-speech segments of the signal.

Consider using (6) as a basis for estimation of the speech variance,  $\varsigma$ . We drop the  $f$  subscript for simplicity. Bayes' theorem gives

$$f(\varsigma | \mathbf{t}, \hat{\nu}) \propto f(\mathbf{t} | \varsigma, \hat{\nu}) f(\varsigma). \quad (7)$$

If we assume a flat prior  $f(\varsigma) \propto 1$ , substituting (6) into (7), differentiating with respect to  $\varsigma$  and equating to zero gives the well known maximum likelihood estimate,

$$\hat{\varsigma} = \max\left(|\mathbf{t}|^2 - \hat{\nu}, 0\right). \quad (8)$$

This is known to provide a “reasonable” estimate of the speech variance, but always requires regularisation. In ASR, the ML estimate is known as power spectral subtraction. It is regularised by means of an over-subtraction factor,  $\alpha$ , and a flooring factor,  $\beta$ :

$$\hat{\varsigma} = \max\left(|\mathbf{t}|^2 - \alpha\hat{\nu}, \beta\hat{\nu}\right). \quad (9)$$

### B. ML SNR estimate

The purpose of the above derivation is to show that a commonly used speech feature can be seen in a Bayesian sense as an estimate of the variance  $\varsigma$ . We now follow the same procedure, but aim from the outset to estimate SNR. Define

$$\xi_f = \frac{\varsigma_f}{\nu_f}, \quad (10)$$

where  $\xi_f$  is exactly the *a-priori* SNR of McAulay and Malpass [9], popularised by Ephraim and Malah [3]. The  $f$  subscript indicates that the SNR is frequency dependent. Substituting  $\varsigma_f = \xi_f\nu_f$  into (6),

$$f(\mathbf{t}_f | \xi_f, \nu_f) = \frac{1}{\pi\nu_f(1 + \xi_f)} \exp\left(-\frac{|\mathbf{t}_f|^2}{\nu_f(1 + \xi_f)}\right). \quad (11)$$

The subscript is dropped again hereafter for simplicity.

This time, the posterior is in terms of  $\xi$ ,

$$f(\xi | \mathbf{t}, \hat{\nu}) \propto f(\mathbf{t} | \xi, \hat{\nu}) f(\xi). \quad (12)$$

Assuming a flat prior, substituting (11) into (12), differentiating and equating to zero,

$$\hat{\xi} = \max\left(\frac{|\mathbf{t}|^2}{\hat{\nu}} - 1, 0\right). \quad (13)$$

It is shown in section IV that this result requires no further normalisation to work well. Further, notice that

$$\log(1 + \hat{\xi}) = \log \left( \max \left[ 1, \frac{|\mathbf{t}|^2}{\hat{\nu}} \right] \right), \quad (14)$$

which is the same form as (3). However, no ad-hoc spectral model is necessary.

We note that in the Decision Directed estimator of [3], the ML estimate of  $\xi$  of (13) is regularised using an estimate based on the previous spectral magnitude estimate. This is further explored by Cohen [10], and is used in a modified form in [4], [11]. Whilst these approaches are beyond the scope of the present study, our approach does not preclude using them.

### C. Marginalisation over noise variance

Thus far we have assumed that an estimate,  $\hat{\nu}$ , of the noise variance is available. The form of (11), however, with multiplicative instead of additive terms in the denominators, allows marginalisation over the noise variance.

If we have  $N$  frames (spectral vectors) of noise,  $\{\mathbf{n}\}_N = \{\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_N\}$ , that are observed in isolation, we can write

$$f(\nu_f | \{\mathbf{n}\}_N) = \frac{\prod_{i=1}^N f(\mathbf{n}_{i,f} | \nu_f) f(\nu_f)}{\int_0^\infty d\nu' \prod_{i=1}^N f(\mathbf{n}_{i,f} | \nu'_f) f(\nu'_f)}, \quad (15)$$

where the products are over the likelihood terms, not the priors. Again, hereafter we drop subscripts for simplicity. The likelihood terms are exactly the form of equation (5), and we arbitrarily choose a non-informative prior  $f(\nu) \propto \nu^{-1}$ . Equation (15) then reduces to the inverse gamma distribution

$$f(\nu | \{\mathbf{n}\}_N) = \frac{B^A}{\Gamma(A)} \nu^{-A-1} \exp\left(-\frac{B}{\nu}\right) \quad (16)$$

where

$$A = N, \quad B = \sum_{i=1}^N |\mathbf{n}_{i,f}|^2. \quad (17)$$

The MAP solution,  $\hat{\nu}$ , of  $\nu$  would be

$$\hat{\nu} = \frac{B}{A+1}, \quad (18)$$

however, we can use the distribution to marginalise over  $\nu$ . Equation (12) becomes

$$f(\xi | \mathbf{t}) \propto f(\xi) \int_0^\infty d\nu f(\mathbf{t} | \xi, \nu) f(\nu | \{\mathbf{n}\}_N). \quad (19)$$

Substituting (11) and (16) into (19), the forms are conjugate and the integral is just the normalising term from the inverse gamma distribution.

$$f(\xi | \mathbf{t}) \propto f(\xi) \times \frac{B^A}{\Gamma(A)} \frac{\Gamma(A+1)}{\xi+1} \left( \frac{|\mathbf{t}|^2 + (\xi+1)B}{\xi+1} \right)^{-(A+1)}. \quad (20)$$

### D. Marginal ML estimate

If we assume a flat prior,  $f(\xi) \propto 1$ , as before, differentiating (20) and equating to zero gives

$$\hat{\xi} = \max \left( \frac{A|\mathbf{t}|^2}{B} - 1, 0 \right) \quad (21)$$

Curiously, equation (21) is basically the same as equation (13).

### E. MAP estimate

Instead of using a flat (improper) prior for the speech variance, it is possible to use a proper prior representing real information. The prior distribution should allow (encourage, even) the SNR to be zero, but should discourage large values; greater than a few tens of decibels. Here we use the gamma distribution

$$f(\xi | \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \xi^{\alpha-1} \exp\left(-\frac{\xi}{\beta}\right). \quad (22)$$

Substituting this into (20), differentiating and equating to zero yields a cubic in  $\xi$

$$a_3 \xi^3 + a_2 \xi^2 + a_1 \xi + a_0 = 0, \quad (23)$$

with

$$\begin{aligned} a_3 &= -1, \\ a_2 &= -\beta + (\alpha - 1)\beta - |\mathbf{t}|^2 / B - 2, \\ a_1 &= -\beta + \beta A |\mathbf{t}|^2 / B + (\alpha - 1)\beta |\mathbf{t}|^2 / B \\ &\quad + 2(\alpha - 1)\beta - |\mathbf{t}|^2 / B - 1, \\ a_0 &= (\alpha - 1)\beta + (\alpha - 1)\beta |\mathbf{t}|^2 / B, \end{aligned} \quad (24)$$

The cubic is readily solved using the cubic equation [12], and always has at least one real root. The root can, however, be negative, so the resulting  $\hat{\xi}$  should be floored at zero.

To set the hyper-parameters, we find that simply constraining the expectation of the gamma distribution to be the average ML SNR of the current frame works satisfactorily,

$$E(\xi_f) = \alpha \beta_f \quad (25)$$

$$\beta_f = \frac{1}{\alpha} E(\xi_f) = \frac{1}{\alpha} \left[ \frac{1}{F} \sum_{f=1}^F \frac{|\mathbf{t}_f|^2}{\hat{\nu}_f} - 1 \right], \quad (26)$$

and, empirically,  $\alpha = 0.01$ .

For illustration, figure 1 shows a histogram of SNR (actually  $|\mathbf{t}|^2 / \hat{\nu}$ ) at 1000 Hz for the clean part of the aurora 2 training data. Also shown is a gamma distribution with  $\alpha = 0.01$  and  $\beta$  set such that the expectation is 48dB, the approximate SNR of the clean aurora 2 data. The plot is in the log domain. Notice that the gamma distribution is basically flat (caused by  $\alpha$  being close to 0), but falls rapidly for high values, i.e., it is largely uninformative but discourages high SNR.

We choose a gamma prior in this study for simplicity. Other authors (a recent example is [13]) have made persuasive cases for the speech prior being closer to a generalised gamma distribution. In ASR, the speech prior is often represented by a large Gaussian mixture [14].

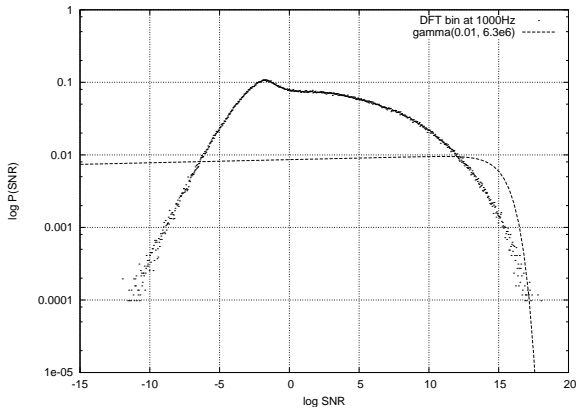


Fig. 1. Histogram of clean data at 1000 Hz and gamma distribution with  $\alpha = 0.01$  and  $\alpha\beta = 48\text{dB}$ . The gamma distribution is largely flat (uninformative), but imposes an upper limit.  $\log(\text{SNR})$  would normally be floored at 0.

#### IV. EXPERIMENTS

To allow comparison with [7] we present experimental results on aurora 2. The aurora 2 task [15] is a well known evaluation for noise compensation techniques. It is a simple digit recognition task with real noise artificially added in 5dB increments such that performance without noise compensation ranges from almost perfect to almost random. Both clean (uncorrupted) and multi-condition (additive noise corrupted) training sets are provided, along with three test sets:

- A Data corrupted with the same noise used in the (multi-condition) training.
- B As test set A, but using different noise.
- C A subset of the noises above, but additionally with a convolutional filter.

Aurora 2 does not distinguish evaluation and test data, so results may be biased towards this dataset and should be considered optimistic.

We used a simple ‘‘MFCC’’ front-end with a 256 point DFT every 10ms. The noise reduction techniques were applied in the power-spectral domain (129 bins), after which a filter bank of 23 mel-spaced triangular bins was applied. The usual logarithm and truncated DCT then produced 13 cepstral coefficients (including C0) plus first and second order delta coefficients. Where CMN and CVN were applied, the means and variances were calculated separately for the whole of each utterance.

The noise values were obtained using the low-energy envelope tracking method described in [5], but with a simplified correction factor from [16]: The 20 lowest energy samples in a sliding 100 frame (1 second) window were averaged, and multiplied by a correction factor,  $C$ . See section V-B for a discussion of this factor.

Complete results are shown in Figure 2. Each graph represents a full aurora evaluation with both multi-condition and clean training. The SNR of clean testing data was measured to be around 48dB, and is off the axis, but the result is shown as the first number in parentheses in the legend. The second

number in the legend is the usual aurora figure of merit: the average of the scores from 0dB to 20dB.

Each graph in the left column represents use of CMN, whereas the right column represents use of CVN (implying CMN also). The four rows are, respectively, the value passed to the filter-bank being

- 1) The usual non-SNR (power spectral) features.
- 2) As 1, but with spectral subtraction.

$$\hat{\xi} = \max\left(|t|^2 - \alpha\hat{\nu}, \beta\hat{\nu}\right), \quad (27)$$

with  $\alpha = 1$  and  $\beta = 0.1$ , found with a coarse grid search.

- 3) One plus the maximum likelihood estimate of SNR from the marginal distribution

$$\hat{\xi} + 1 = \max\left(\frac{A|t|^2}{B} - 1, 0\right) + 1, \quad (28)$$

$$= \max\left(\frac{A|t|^2}{B}, 1\right). \quad (29)$$

- 4) One plus the MAP estimate of the SNR with a gamma prior,

$$\hat{\xi} + 1, \quad (30)$$

where  $\hat{\xi}$  is the solution of the cubic in (23) and (24).

We stress that these results are not state of the art for this database; the purpose is to compare techniques.

#### V. DISCUSSION

##### A. Performance

The most significant result of these experiments is that the CVN results for the SNR features agree with, even exceed, those in [7]. This is despite the fact that no involved spectral model is used to distinguish the speech and noise. It seems that simply being able to track the background noise level with the low-energy envelope is enough.

The use of the simple gamma prior has a small benefit, but at the cost of an extra parameter and finding the solution to a cubic equation. Whilst this is not computationally onerous, it is doubtful whether it is worthwhile given the good performance of the much simpler ML solution. However, the spirit of the approach is important; it shows a principled way to incorporate prior information.

Spectral subtraction gives an improvement over the baseline, but does not respond to CVN. This is at odds with the results in [8], but in agreement with our own anecdotal evidence. This is a curious result since there is not a large theoretical difference between SNR features and spectral subtraction. The practical difference between the two is that SNR features normalise before the filter-bank, whereas CMN works after it. If we denote the filter-bank weights for a single bin by  $w_1, w_2, \dots$ , the SNR features presented to the decoder are of the form

$$\log(1 + w_1\xi_1 + w_2\xi_2 + \dots), \quad (31)$$

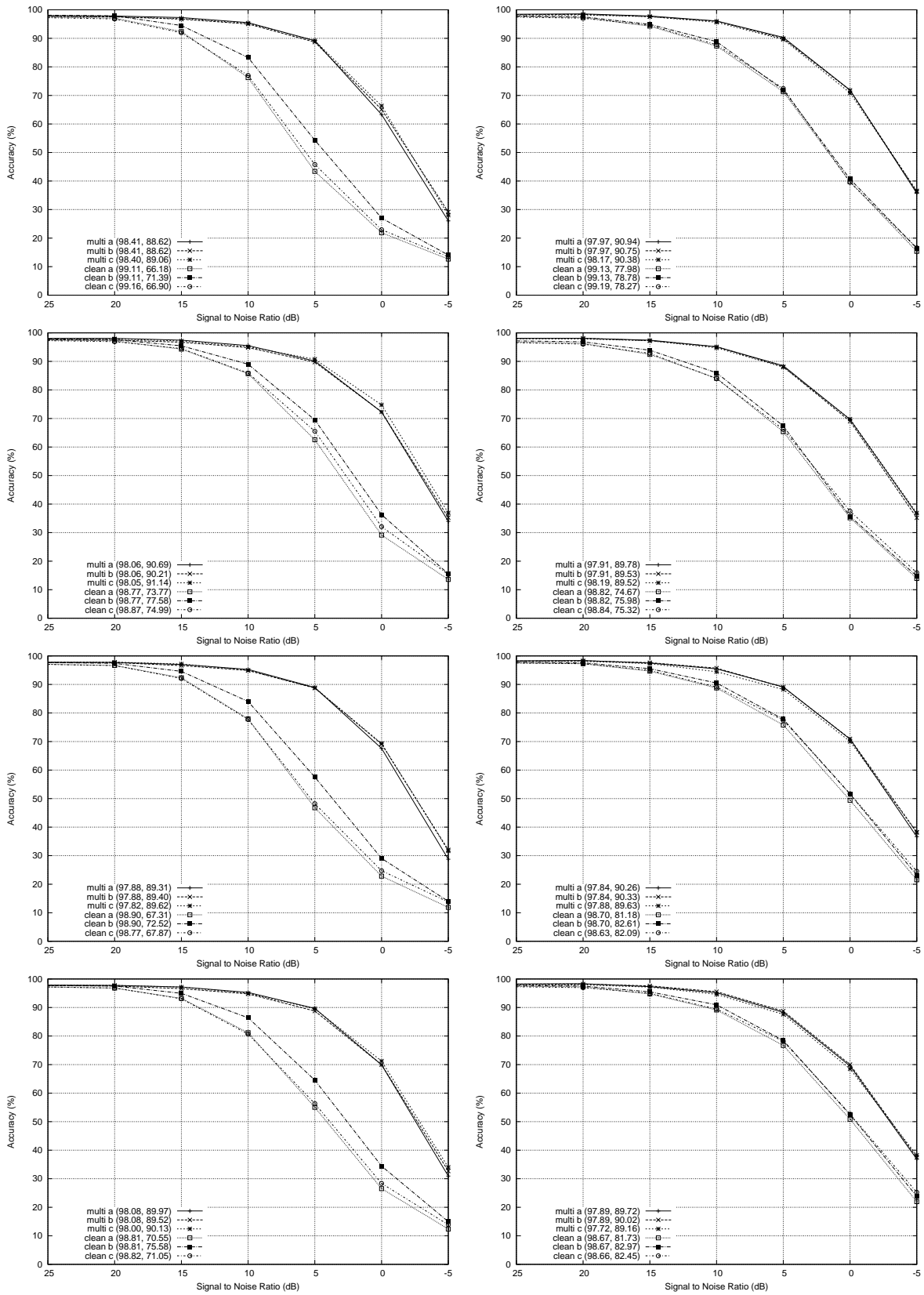


Fig. 2. Results. The left column is for CMN, right for CMN+CVN. The four rows top to bottom are respectively: No noise compensation, Spectral Subtraction, ML SNR and MAP SNR with gamma prior. See the text for details of the database.

whereas the spectral subtraction features are *closer* to the form

$$\log \left( 1 + \frac{w_1 s_1 + w_2 s_2 + \dots}{w_1 n_1 + w_2 n_2 + \dots} \right). \quad (32)$$

Given that  $\log(a + b) \approx \log \max(a, b)$ , we hypothesise that a large noise component anywhere in the band spanned by the given bin could dominate the latter expression. This in turn offers some explanation for the improved performance of SNR features. It remains a subject for further investigation.

### B. Noise tracker correction factor

The low-energy envelope tracker normally requires correction as its estimate is biased too small. In [16], Lathoud *et al.* suggest that a multiplicative correction factor

$$C = \frac{1}{(1.5\gamma)^2}, \quad (33)$$

works well, where  $\gamma$  is the fraction of samples assumed to be noise. In our case,  $\gamma = 0.2$  so  $C = 11.1$ . In fact, we found that, whilst this correction factor was necessary for the spectral subtraction approach, a value of  $C = 1$  was better for SNR features (the results in Figure 2 are for these values).

It is tempting to conclude that SNR features do not need a correction factor. However, it is more likely that the noise tracker with  $C = 1$  was producing noise estimates about 11 times too small, so the SNR estimates were 11 times too large. Writing the situation as

$$\log(1 + 11\xi) = \log(11) + \log \left( \frac{1}{11} + \xi \right), \quad (34)$$

it is clear that this corresponds to using a smaller floor in the logarithm. This floor is also very close to the one empirically found to work well as the parameter  $\beta$  in spectral subtraction.

The low-energy envelope is a noise floor rather than a noise estimate; it is intuitively reasonable that this floor is also the level below which speech and noise cannot be distinguished. We hypothesise that the optimal value of  $C$  in low-energy envelope tracking is the same as the optimal floor for SNR. Thus, when using SNR based features, these values cancel out giving a parameter-free feature. Proof that  $C = 1$  is optimal for SNR features, however, will require a careful mathematical and experimental analysis.

## VI. CONCLUSIONS

SNR features for ASR have several practical and mathematical advantages over the more usual spectral power features. The naive SNR estimate is actually the optimal estimate under a fairly rigorous Bayesian analysis, and the framework leaves room for further incorporation of prior information, as is common recently in ASR. SNR features perform well in noisy conditions, and outperform other features when combined with CVN. Prior information incorporated via a gamma prior distribution improves results still further, although the difference may not merit the extra complexity. In practice a different prior form, or one trained on real data ought to work better.

We have some evidence that the optimal correction factor used in low-energy envelope tracking cancels exactly the

flooring used in the logarithm for SNR features, making SNR features almost parameter-free when noise is estimated in this manner.

## VII. ACKNOWLEDGEMENTS

The author is grateful to Mathew Magimai-Doss for comments on the manuscript. This work was supported by the Swiss National Center of Competence in Research on Interactive Multi-modal Information Management. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

## REFERENCES

- [1] O. Viikki and K. Laurila, "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization," in *Robust Speech Recognition for Unknown Communication Channels*. ISCA, April 1997, pp. 107–110, Pont-à-Mousson, France.
- [2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, pp. 113–120, April 1979.
- [3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.
- [4] ETSI, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," ETSI, ETSI Standard 202 050, 2002, v1.1.1.
- [5] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Communication*, no. 34, pp. 141–158, 2001.
- [6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.
- [7] G. Lathoud, M. Magimai-Doss, B. Mesot, and H. Bourlard, "Unsupervised spectral subtraction for noise-robust ASR," in *Proceedings of the 2005 IEEE ASRU Workshop*, December 2005, San Juan, Puerto Rico.
- [8] J. C. Segura, M. C. Benítez, A. de la Torre, and A. J. Rubio, "Feature extraction combining spectral noise reduction and cepstral histogram equalisation for robust ASR," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 225–228.
- [9] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 2, pp. 137–145, April 1980.
- [10] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, September 2005.
- [11] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, "A two-step noise reduction technique," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, Montreal, Canada, May 2004, pp. 289–292.
- [12] "Cubic function," Wikipedia. [Online]. Available: [http://en.wikipedia.org/wiki/Cubic\\_function](http://en.wikipedia.org/wiki/Cubic_function)
- [13] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1741–1752, August 2007.
- [14] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Atlanta, US, May 1996, pp. 733–736.
- [15] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, September 2000, Paris, France.
- [16] G. Lathoud, M. Magimai-Doss, and H. Bourlard, "Channel normalization for unsupervised spectral subtraction," Idiap Research Institute, IDIAP-RR 06-09, February 2006. [Online]. Available: <http://publications.idiap.ch>



# Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition

Philip N. Garner\*

*Idiap Research Institute, Centre du Parc, Rue Marconi 19, P.O. Box 592, 1920 Martigny, Switzerland*

Received 29 December 2010; received in revised form 16 May 2011; accepted 16 May 2011

Available online 26 May 2011

## Abstract

Cepstral normalisation in automatic speech recognition is investigated in the context of robustness to additive noise. In this paper, it is argued that such normalisation leads naturally to a speech feature based on signal to noise ratio rather than absolute energy (or power). Explicit calculation of this *SNR-cepstrum* by means of a noise estimate is shown to have theoretical and practical advantages over the usual (energy based) cepstrum. The relationship between the SNR-cepstrum and the articulation index, known in psycho-acoustics, is discussed. Experiments are presented suggesting that the combination of the SNR-cepstrum with the well known perceptual linear prediction method can be beneficial in noisy environments.

© 2011 Elsevier B.V. All rights reserved.

*Keywords:* Automatic speech recognition; Cepstral normalisation; Noise robustness; Aurora

## 1. Introduction

An important problem encountered in speech signal processing is that of how to normalise a signal for the effects of noise. In speech enhancement the task is to remove noise from a signal to reproduce the uncorrupted signal such that it is perceived by a listener to be less noisy. In automatic speech recognition (ASR), the task is to reduce the effect of noise on recognition accuracy. This paper concentrates on the latter (ASR) problem.

Two categories of noise are generally considered: Additive noise is that which represents a distinct signal other than the one of interest. Convolutional noise is that which alters the spectral shape, and can be associated with either the signal of interest, or both the signal and the additive noise.

The present work stems from the practical experience that it is very difficult to improve upon cepstral normalisation techniques for noise robustness. Cepstral mean nor-

malisation (CMN) (Furui, 1981) is a well established technique that compensates, in a theoretically sound way, for convolutional noise. It is based on the persuasive observation that a linear channel distortion becomes a constant offset in the cepstral domain. More heuristically, CMN also affords some robustness to additive noise. Cepstral variance normalisation (CVN) (Viikki and Laurila, 1997; Viikki and Laurila, 1998) generally results in very good noise robustness, but the reason for this is not well understood.

Many common practical solutions for additive noise compensation are based on the assumption of a simple additive Gaussian model for both speech and noise in the spectral domain. In ASR, the spectral subtraction approach of Boll (1979) is well established. In speech enhancement, much work is based on the technique of Ephraim et al. (1984). Both these techniques have influenced the design of the ETSI (2002) standard ASR front-end. However, at least in a batch mode of operation, and certainly combined with multi-condition training, CMN combined with CVN can exceed the performance of all these techniques.

\* Tel.: +41 27 721 7768.

E-mail address: [pgarner@idiap.ch](mailto:pgarner@idiap.ch)

In this paper, building on previous work, the theoretical effect of CMN and CVN in additive noise is studied. It is shown that the use of CMN implies that the features presented to an ASR decoder are in fact measures of (log) signal to noise ratio (SNR) rather than (log) energy. Based on this observation, a SNR feature is derived formally, the derivation providing both theoretical and practical advantages over the equivalent for energy based features.

The SNR-cepstrum is then placed in context amongst other techniques, emphasising that there is a great deal of commonality between noise robustness in ASR, speech enhancement and indeed the workings of the inner ear.

The paper is split roughly into two parts. Sections 1–4 are largely theoretical, expanding previous work to give a thorough basis for the SNR-cepstrum. Sections 5–7 proceed to evaluate the SNR-cepstrum in the context of the linear predictive features that are common in modern ASR systems.

## 2. Background

In a simplistic, but informative, view of an ASR front-end, an acoustic signal is Fourier transformed to give a vector of spectral coefficients  $(s_1, s_2, \dots, s_F)^T$ . After a linear transform (filter-bank) implementing a non-linear frequency warp, the cepstrum is calculated. The cepstrum involves a logarithm followed by another linear transform (DCT).

### 2.1. Convolutional noise

Although only one is normally considered, note that two types of convolutional noise can be distinguished:

1. A *source* noise,  $\mathbf{g} = (g_1, g_2, \dots, g_F)^T$ , associated only with the speech signal. This can be thought of as being representative of a speaker.
2. A *channel* noise,  $\mathbf{h} = (h_1, h_2, \dots, h_F)^T$ , associated with the microphone and transmission channel.

In the presence of convolutional noise, which is multiplicative in the frequency domain, the logarithm for each frequency bin,  $f \in \{1, 2, \dots, F\}$ , becomes

$$\log(h_f g_f s_f) = \log(h_f) + \log(g_f) + \log(s_f), \quad (1)$$

where  $\log(s_f)$  varies, and  $\log(h_f)$  is constant over time.  $\log(g_f)$  is taken to represent the component of the speech that is constant over time, being some characteristic of the speaker.

It follows from Eq. (1) that, if  $\log(s_f)$  can be assumed to have zero mean, the noise terms can be removed by subtracting the long term average of the log-spectrum. This is achieved by cepstral mean normalisation (Furui, 1981, although the technique has been attributed to Atal even earlier) or by the RASTA processing of Hermansky and Morgan (1994). Note also that, when the filter-bank is con-

sidered, the above holds if the  $h_f$  and  $g_f$  are assumed constant within a given filter-bank bin.

### 2.2. Additive noise

When additive noise is also present, typically it is assumed to remain additive after the Fourier transform. In this sense, the logarithm operation becomes

$$\log(h_f g_f s_f + h_f n_f) = \log(h_f) + \log(g_f s_f + n_f). \quad (2)$$

where  $(n_1, n_2, \dots, n_F)^T$  is the noise spectrum. From Eq. (2), it appears that CMN and the like cannot work in significant additive noise unless the additive noise is removed first. To this end, there is a large body of work focusing on additive noise removal. In ASR, the spectral subtraction approach of Boll (1979) was further developed by, for instance, Van Compernelle (1989), and is well established. It is often used as a means to derive a Wiener filter. In speech enhancement, much work is based on the technique of Ephraim et al. (1984).

The state of the art in additive noise robustness is probably in the body of work based on the additive model of Acero and Stern (1990), and the vector Taylor series approach of Moreno et al. (1996). Such techniques are characterised by a large Gaussian mixture prior on the speech signal, a recent exemplar being Li et al. (2007). It is not the goal of the present paper to approach the performance of such techniques. Rather, a building block is presented that could be used in combination with these techniques.

### 2.3. SNR features

The logarithm of a sum can be written

$$\begin{aligned} \log(x + a) &= \log(a) + \frac{x}{a} - \frac{x^2}{2a^2} + \frac{x^3}{3a^3} \dots \\ &= \log(a) + \log\left(1 + \frac{x}{a}\right). \end{aligned} \quad (3)$$

Although the relationship is clear without the series expansion, the latter emphasises that the term  $\log(a)$  is the component that is independent of  $x$ . This in turn suggests that Eq. (2) might better be written

$$\log(h_f g_f s_f + h_f n_f) = \log(h_f n_f) + \log\left(1 + \frac{g_f s_f}{n_f}\right), \quad (4)$$

emphasising that CMN would actually remove the constant term  $\log(h_f n_f)$ , or its mean if either  $h_f$  or  $n_f$  were non-deterministic.

It appears from the above analysis that, if CMN is used, the features that are presented to the ASR decoder are actually (a linear transform of) the logarithm of one plus the signal to noise ratio (SNR). This will happen even if the additive noise is simply the minimal background noise usually associated with clean recordings. It follows that one could try to calculate the SNR from the outset rather than calculate a spectral power measure and rely on CMN to

produce the SNR. A-priori, such an approach has at least three appealing properties:

1. The flooring of the logarithm happens naturally. The SNR (expressed as a power ratio) cannot fall below zero, so the argument of the logarithm is naturally floored at unity, and the logarithm is hence positive.
2. SNR is inherently independent of  $\mathbf{h}$ , the convolutional noise associated with microphones and the gain associated with pre-amplifiers.
3. If applied before the filter bank, the assumption that  $h_f$  remains constant over the range of the filter bin is no longer required.

It turns out that SNR is also mathematically appealing.

Notice that, whilst the channel noise,  $h_f$ , is cancelled by taking the SNR, the source noise,  $g_f$ , is still present. However, for high SNR it will be removed by CMN. It follows that the SNR is not a replacement for CMN in its speaker normalisation sense. It also suggests that direct comparison of SNR based features with CMN would not be fair.

### 3. The SNR spectrum

In contrast to the previous section, which was left deliberately simplistic, a more rigorous derivation of a SNR based feature is now presented. After defining a Gaussian model of speech in noise, the derivation proceeds by showing that power spectral subtraction can be seen as a particular maximum-likelihood (ML) solution. Two ML estimators for the SNR are then derived.

#### 3.1. Gaussian model

Assume that a DFT operation produces a vector,  $\mathbf{x}$ , with complex components,  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_F$ , where the real and imaginary parts of each  $\mathbf{x}_f$  are Gaussian, independent and identically distributed (i.i.d.) with zero mean and variance  $v_f$ . That is,

$$p(\mathbf{x}_f|v_f) = \frac{1}{\pi v_f} \exp\left(-\frac{|\mathbf{x}_f|^2}{v_f}\right). \quad (5)$$

In the case where two coloured noise signals are distinguished, a background noise,  $\mathbf{n}$ , and a signal of interest,  $\mathbf{s}$ , typically speech, denote the noise variance as  $\mathbf{v}$  and the speech variance as  $\zeta$ . In general, the background noise can be observed in isolation and modelled as

$$p(\mathbf{n}_f|v_f) = \frac{1}{\pi v_f} \exp\left(-\frac{|\mathbf{n}_f|^2}{v_f}\right). \quad (6)$$

The speech, however, cannot normally be observed in isolation. It is always added to noise. When both speech and additive noise are present the variances add, meaning that the total signal,  $\mathbf{t}_f = \mathbf{s}_f + \mathbf{n}_f$ , can be modelled as

$$p(\mathbf{t}_f|\zeta_f, v_f) = \frac{1}{\pi(\zeta_f + v_f)} \exp\left(-\frac{|\mathbf{t}_f|^2}{\zeta_f + v_f}\right). \quad (7)$$

Although neither the Gaussian nor i.i.d. assumptions are likely to be true in practice, the above model is the basis of the Wiener filter and of the widely used Ephraim et al. (1984) speech enhancement technique. The goal is usually formulated as requiring an estimate of  $\mathbf{s}_f$ . However, it is first necessary to find an estimate of  $\zeta_f$ .

#### 3.2. Variance as an ASR feature

The well known maximum likelihood estimate of  $\zeta_f$  is instructive in determining the right approach for the definition and estimation of SNR. It proceeds as follows, where the  $f$  subscript is dropped for simplicity: Assume that an estimate,  $\hat{\mathbf{v}}$ , of  $\mathbf{v}$  is available via solution of (6) during, for instance, non-speech segments of the signal. The estimate of the speech variance,  $\zeta$ , then follows from Bayes' theorem,

$$p(\zeta|\mathbf{t}, \hat{\mathbf{v}}) \propto p(\mathbf{t}|\zeta, \hat{\mathbf{v}})p(\zeta|\hat{\mathbf{v}}). \quad (8)$$

Assuming  $p(\zeta|\hat{\mathbf{v}}) = p(\zeta)p(\hat{\mathbf{v}})$  and a flat prior  $p(\zeta) \propto 1$ , substituting (7) into (8), differentiating with respect to  $\zeta$  and equating to zero gives the ML estimate,

$$\hat{\zeta} = \max\left(|\mathbf{t}|^2 - \hat{\mathbf{v}}, 0\right). \quad (9)$$

Notice that, in ASR at least, this is simply power spectral subtraction. More generally, it is known to provide a “reasonable” estimate of the speech variance, but always requires regularisation. In ASR, it is regularised by means of an over-subtraction factor,  $\alpha$ , and a flooring factor,  $\beta$ :

$$\hat{\zeta} = \max\left(|\mathbf{t}|^2 - \alpha\hat{\mathbf{v}}, \beta\hat{\mathbf{v}}\right), \quad (10)$$

as in (Van Compernelle, 1989).

The above derivation shows that a commonly used speech feature can be seen in a Bayesian sense as an estimate of the variance  $\zeta$ . This interpretation is reinforced when convolutional noise is considered. Making the substitution  $\mathbf{\eta}_f = \sqrt{h_f}\mathbf{x}_f$  in Eq. (5), the Jacobian determinant is  $h_f^{-1}$ , so

$$p(\mathbf{\eta}_f|h_f, v_f) = \frac{1}{\pi h_f v_f} \exp\left(-\frac{|\mathbf{\eta}_f|^2}{h_f v_f}\right), \quad (11)$$

i.e., the convolutional term multiplies the variance, exactly as in the simplistic model of Section 2.

The above implies that estimation for the purposes of ASR can focus on the variance,  $\zeta$ , rather than the (uncorrupted) observation,  $\mathbf{s}$ , as in enhancement.

#### 3.3. ML SNR estimate

Motivated by the term of interest being the variance, define the SNR as

$$\xi_f = \frac{\zeta_f}{v_f}, \quad (12)$$

The  $f$  subscript indicates that the SNR is frequency dependent. Substituting  $\zeta_f = \xi_f v_f$  into (7),

$$p(t_f | \xi_f, v_f) = \frac{1}{\pi v_f (1 + \xi_f)} \exp\left(-\frac{|t_f|^2}{v_f (1 + \xi_f)}\right). \quad (13)$$

The subscript is dropped again hereafter for simplicity.

This time, the posterior is in terms of  $\xi$ ,

$$p(\xi | t, \hat{v}) \propto p(t | \xi, \hat{v}) p(\xi | \hat{v}). \quad (14)$$

Assuming a flat prior, substituting (13) into (14), differentiating and equating to zero,

$$\hat{\xi} = \max\left(\frac{|t|^2}{\hat{v}} - 1, 0\right). \quad (15)$$

### 3.4. Marginalisation over noise variance

Thus far it has been assumed that an estimate,  $\hat{v}$ , of the noise variance is available. In a Bayesian sense, however, the noise is a nuisance variable, the correct approach being to marginalise over it. In the case of variance estimation, such marginalisation is not easily tractable. By contrast, the form of (13), with multiplicative instead of additive terms in the denominators, presents no major difficulty for marginalisation.

If there are  $N$  frames (spectral vectors) of noise,  $\{\mathbf{n}\}_N = \{\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_N\}$ , that are observed in isolation, one can write

$$p(v_f | \{\mathbf{n}\}_N) = \frac{\prod_{i=1}^N p(\mathbf{n}_{i,f} | v_f) p(v_f)}{\int_0^\infty dv' \prod_{i=1}^N p(\mathbf{n}_{i,f} | v') p(v')}, \quad (16)$$

where the products are over the likelihood terms, not the priors. Again, hereafter subscripts are dropped for simplicity. The likelihood terms are exactly the form of Eq. (6), and a non-informative prior,  $p(v) \propto v^{-1}$ , is arbitrarily chosen. Eq. (16) then reduces to the inverse gamma distribution

$$p(v | \{\mathbf{n}\}_N) = \frac{B^A}{\Gamma(A)} v^{-A-1} \exp\left(-\frac{B}{v}\right) \quad (17)$$

where

$$A = N, \quad B = \sum_{i=1}^N |\mathbf{n}_{i,f}|^2. \quad (18)$$

The MAP solution,  $\hat{v}$ , of  $v$  would be

$$\hat{v} = \frac{B}{A+1}, \quad (19)$$

however, the distribution can be used to marginalise over  $v$ . Assuming the prior on SNR is independent of the noise estimate, Eq. (14) becomes

$$p(\xi | t) \propto p(\xi) \int_0^\infty dv p(t | \xi, v) p(v | \{\mathbf{n}\}_N). \quad (20)$$

Substituting (13) and (17) into (20), the forms are conjugate and the integral is just the normalising term from the inverse gamma distribution.

$$p(\xi | t) \propto p(\xi) \times \frac{B^A}{\Gamma(A)} \frac{\Gamma(A+1)}{\xi+1} \left(\frac{|t|^2 + (\xi+1)B}{\xi+1}\right)^{-(A+1)}. \quad (21)$$

If a flat prior,  $p(\xi) \propto 1$ , is assumed as before, differentiating (21) and equating to zero gives a marginal ML estimate:

$$\hat{\xi} = \max\left(\frac{A|t|^2}{B} - 1, 0\right) \quad (22)$$

Curiously, Eq. (22) is basically the same as Eq. (15). It was shown by Garner (2009) that this result requires no further regularisation to work well.

Hereafter, the SNR vector,  $\xi$ , is referred to as the SNR-spectrum. This leads to the resulting cepstrum being called the SNR-cepstrum.

## 4. Context

Whilst the above derivation is novel to the knowledge of the author, the SNR-spectrum is by no means a new concept. Rather, it draws together several loosely related topics.

### 4.1. Enhancement

$\xi$  is exactly the *a-priori* SNR of McAulay et al. (1980), popularised by Ephraim et al. (1984). In enhancement, this measure is used as an intermediate result in the reconstruction of an enhanced spectrum. The Wiener filter can be defined in terms of the SNR:

$$w = \frac{\xi}{\xi+1}. \quad (23)$$

In the decision directed estimator of Ephraim et al. (1984), the ML estimate of  $\xi$  of (15) is regularised using an estimate based on the previous spectral magnitude estimate. This is further explored by Cohen (2005), and is used in a modified form in ETSI (2002), Plapous et al. (2004). Whilst these approaches are beyond the scope of the present study, the proposed approach does not preclude using them.

### 4.2. Automatic speech recognition

Lathoud et al. (2005) present an ad-hoc model allowing a signal to be described in terms of noise and speech spectra. Those authors perform what they refer to as “Unsupervised” spectral subtraction. In fact, they explicitly floor the SNR using (in the present notation)

$$\max\left(1, \frac{s_f}{n_f}\right). \quad (24)$$

Notice that

$$\log(1 + \hat{\xi}) = \log\left(\max\left[1, \frac{|t|^2}{\hat{v}}\right]\right), \quad (25)$$

which is the same form as (24). However, no ad-hoc spectral model is necessary. It was shown by Garner (2009) that this formulation can actually exceed the performance reported by Lathoud et al. (2005).

The terminology raises an interesting issue: in the context of CMN, there is little difference between using the SNR-spectrum, and spectral subtraction. This is explored below in Section 6.2.

#### 4.3. Relationship with articulation index

Allen (1994) describes earlier work by Fletcher analysing the probable workings of the inner ear. In particular, Allen states that Fletcher's experiments suggest that the cochlea is sensitive to SNR:

The signal to noise ratio of each cochlear inner hair cell signal is important to the formation of the feature channels since [the channel error] is known to depend directly on these SNRs rather than on spectral energy.

Later, Allen (2005) defines the articulation index (AI) as

$$AI_k = \min\left(\frac{1}{3}\log_{10}(1 + c^2\text{snr}_k^2), 1\right). \quad (26)$$

The AI is lower bounded at 0 by the logarithm, and upper bounded at 1 by a heuristic 30 dB dynamic range of speech.

Notice that the AI has the same form, except for linear transformation, as the speech feature described above that arises from CMN. This in turn is known to work well in ASR. These two derivations are totally independent. It follows that, under CMN, the feature being presented to an ASR decoder is the AI, just as in the human ear.

In fact, the AI has been used directly as an ASR feature by Lobdell et al. (2008). The approach of those authors was to use the AI specifically to mimic the function of the ear. In this sense, the present approach is complementary, driven more mathematically than perceptually.

#### 4.4. Noise tracker

In order to obtain a noise estimate, Garner (2009) used the low-energy envelope tracker advocated by Lathoud et al. (2006), based on Ris and Dupont (2001) and Martin (2001). The low-energy envelope tracker normally requires correction as its estimate is biased too small. Lathoud et al. (2006) suggest that a multiplicative correction factor

$$C = \frac{1}{(1.5\gamma)^2}, \quad (27)$$

works well, where  $\gamma$  is the fraction of samples assumed to be noise. However, Garner (2009) found that a value of  $C = 1$  was better for the SNR-cepstrum, rather than the  $C \approx 11$

that would be implied from Eq. (27). This in turn implied that the feature being presented to the decoder was closer to

$$\log(1 + 11\xi) = \log(11) + \log\left(\frac{1}{11} + \xi\right). \quad (28)$$

The right hand side of Eq. (28) implies that this corresponds to using a smaller floor in the logarithm. Further, it is close to the one empirically found to work well as the parameter  $\beta$  in spectral subtraction. However, the left hand side of Eq. (28) suggests a relationship with the AI: Allen (2005) states that the value  $c$  from Eq. (26) should be around 2. The square is certainly the same order of magnitude as the 11 that occurs empirically in the results of Garner (2009).  $C$  is based on noise minima and  $c$  is based on speech maxima; whatever the actual value of these constants, the present approach is unable to distinguish them. However, that they appear to cancel each other out suggests they have the same origin.

#### 4.5. Cepstral variance normalisation

Whilst cepstral variance normalisation (CVN) is known to provide noise robustness (Viikki and Laurila, 1997, 1998), the justification for this is normally attributed to a heuristic and brute force shift of the observation PDF towards that of the model. This heuristic is used to good advantage in histogram normalisation (Segura et al., 2002; de la Torre et al., 2005). In the context of the SNR-spectrum, however, the concept of CVN is far more tangible: it is normalising SNR dynamic range.

As an aside, it follows that it may be possible to normalise for SNR at some other point in the processing chain. This has been investigated by the author without success. An obvious tentative conclusion is that the removal of the source noise,  $\mathbf{g}$ , via CMN is important beforehand.

#### 4.6. Summary

The SNR-spectrum arises as a natural consequence of doing CMN on ASR features. CVN then takes on a physical interpretation as normalisation of the SNR dynamic range in dB. If defined more formally as the ratio of speech and noise variances, the intuitive estimator of SNR is also the marginal ML estimator under Gaussian noise.

The SNR-cepstrum appears to be exactly (differing only by linear transform) the AI of Fletcher as defined by Allen, suggesting a close relationship with the sensory mechanisms in the cochlea. Calculating the SNR-cepstrum as suggested both by the cochlea and practical computation leads to better noise robustness at low SNR.

## 5. Experiments

### 5.1. Previous results

Garner (2009) presented results showing that SNR based MFCC (mel frequency cepstral coefficients) features

were more noise robust than the usual energy based features on the aurora 2 database. The aurora 2 task (Hirsch and Pearce, 2000) is a well known evaluation for noise compensation techniques. It is a simple English digit recognition task with real noise artificially added in 5 dB increments such that performance without noise compensation ranges from almost perfect to almost random. Both clean (uncorrupted) and multi-condition (additive noise corrupted) training sets are provided, along with three test sets:

- A Data corrupted with the same noise used in the (multi-condition) training.
- B As test set A, but using different noise.
- C A subset of the noises above, but additionally with a convolutional filter.

Aurora 2 does not distinguish evaluation and test data, so results may be biased towards this data-set and should be considered optimistic. It should also be stressed that the results in this paper are not state of the art for this database; the purpose is to compare techniques.

Aurora 2 is very useful for optimisation and evaluation of front-ends; this is because it runs quickly and has a thorough test set. However, several criticisms can be levelled at aurora 2:

1. It is real noise, but added artificially. This assumes that the additive noise assumption is exact, and ignores effects associated with the fact that speakers will modify their voices to compensate for noise presence.
2. It is digits, hence with a limited grammar and incomplete phonetic coverage.

There is also a somewhat intangible feeling in the community that aurora 2 results are often not reflected in real world systems.

The results of Garner (2009) are summarised in Fig. 1. Each graph represents a full aurora 2 evaluation for either multi-condition or clean training. As the results for the different test sets (A, B and C) are virtually indistinguishable when CMN is used, each curve is the average of the three sets. The SNR of clean testing data was measured to be around 48 dB, and is off the axis, but the result is shown as the first number in parentheses in the legend. The second number in the legend is the usual aurora 2 figure of merit: the average of the scores from 0 dB to 20 dB. Both numbers are averaged over the three test sets.

The first curve in Fig. 1 shows an MFCC baseline using CMN in clean (mismatched) training conditions. The following two curves show the benefits of using CVN too (CMVN: cepstral mean and variance normalisation), and of multi-condition (matched) training. The next curve shows that spectral subtraction cannot improve on CVN, whilst the penultimate curve shows that the SNR-cepstrum can further improve on CVN in mismatched conditions. The final curve shows that the SNR-cepstrum does not

afford any further improvement in matched conditions. In fact, all techniques perform very similarly under multi-condition training.

Notice that, whilst the aurora 2 figure of merit is higher for the SNR based features, it is mainly gained from improvements below about 15 dB SNR. In cleaner conditions, the usual energy based features perform better. It seems reasonable to attribute this difference to the noise tracker. Certainly the noise tracker is imperfect, and it is the only major difference between the two techniques at high SNR.

## 5.2. Hypotheses

In the present investigation, two hypotheses are under test:

1. State of the art systems often use linear prediction features as alternatives to the MFCCs used in previous work. Do such features also benefit from the use of SNR based features?
2. The previous experiments were limited to the scope of aurora 2. Do the benefits of SNR based features transcend the restrictions of this database?

## 5.3. Perceptual linear prediction

Linear prediction (LP) is a common speech analysis method that represents speech using an all pole model (Makhoul, 1975). In the context of ASR, it is used to smooth a spectrum based on the fact that the signal originates from a vocal tract.

LP is normally used in ASR in the form of the perceptual linear prediction (PLP) of Hermansky (1990). PLP modifies the auto-correlation calculation in the first stage of the LP calculation as follows:

1. The power spectrum is binned into critical bands separated according to the bark scale.
2. The bands are weighted according to an equal loudness criterion.
3. The bands are compressed by a cube root representing the power law of human hearing.

PLP has become quite widely used in state of the art ASR systems, e.g., the AMIDA system of Hain et al. (2009). In this sense, it merits investigation in the SNR-spectrum framework.

Whilst LP has a rigorous mathematical underpinning, PLP is more a set of heuristics. That is, the spectral warping is not derived as such, it is introduced in an ad-hoc, but intuitively reasonable manner. Using the same intuition, PLP cepstra can be calculated based on SNR rather than energy. If PLP is seen as simply a smoothing operation, it is reasonable to assume that the same smoothing can

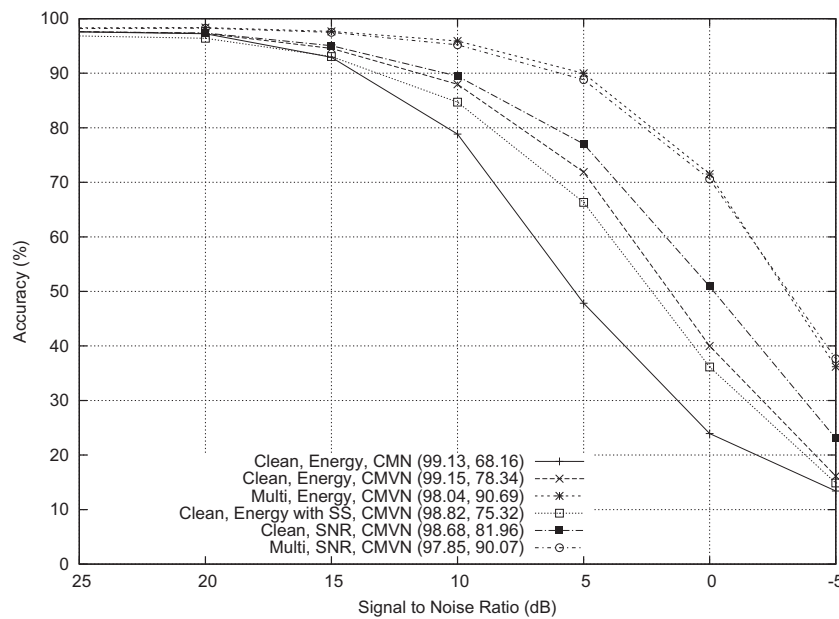


Fig. 1. A summary of previous aurora 2 results for MFCC features. See the text for a description.

be applied to the SNR spectrum rather than the power (energy) spectrum.

#### 5.4. Method

Features in the spirit of PLP were extracted using the Tracter toolkit (Garner and Dines, 2010). That is, pre-emphasis was used in lieu of an equal loudness weighting, then a 256 point DFT was performed every 10ms. The power spectrum of 129 bins was applied to a filter bank of 32 mel-spaced triangular bins (rather than bark spaced trapezoidal bins). The filter bank was cube root compressed (initially), then the usual DCT and LP recursions yielded 13 cepstral coefficients (including C0) plus first and second order delta coefficients. Cepstral means and variances were calculated separately for the whole of each utterance; all new results in this paper use both CMN and CVN.

The SNR based PLP features were extracted as above, except using one plus the ML estimate of the SNR as described in Section 3.4. The LP calculation was as above, except that no cube root compression was employed. This was found to improve performance significantly, and is discussed later in Section 6.3.

Following Garner (2009), the noise values were obtained using the low-energy envelope tracking method described by Ris and Dupont (2001), but with a simplified correction factor of Lathoud et al. (2006): The 20 lowest energy samples in a sliding 100 frame (1 second) window were averaged, but not multiplied by any correction factor.

#### 5.5. Aurora 2 results

Results are shown in Fig. 2. The energy based PLP features perform similarly to the energy based MFCC

features. However, the improvement for SNR based features is considerably more than that for MFCCs in the mismatched (clean training) case. This is encouraging; it strongly suggests not only that the SNR spectrum is applicable to PLP features, but that it is more suited to PLP features than to MFCCs.

#### 5.6. Aurora 3 and 4 results

Aurora 3 and 4 go some way to combat the criticisms that are often levelled at aurora 2.

Aurora 3 is a digit subset of SpeechDat-Car; that is, a similar task to aurora 2 but uttered in real noise. The noise is various driving conditions of a car. Several languages are available; the present experiments are performed on the German (Netsch, 2001) and Danish (Lindberg, 2001) versions. As with aurora 2, a standardised train and test harness is provided using HTK. However, as the noise conditions are real, only three conditions are defined:

wm is *well-matched*; a mixture of all conditions and microphones for both training and testing.

mm is *mid-mismatch*; training with quiet and low noise data on a hands free microphone, testing on high noise data from the same microphone.

hm is *high-mismatch*; training in all conditions on a close talking microphone, testing in low and high noise on a hands free microphone.

No SNR information is immediately available for the Danish database. However, Netsch (2001) gives SNR distributions for the various microphones and conditions. The close talking microphone averages around 20 dB, and the hands free microphones averages around 5–10 dB; however all conditions spread 10 dB either side of

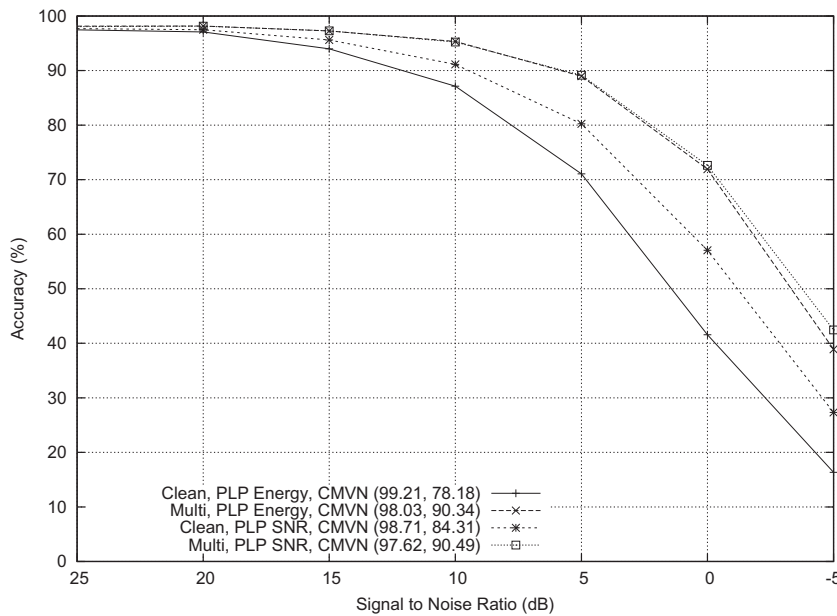


Fig. 2. PLP results on aurora 2 database.

the average. Given these broad measurements, and comparing with aurora 2 results, a-priori it may be expected that SNR features may not afford any improvement on the wm and mm conditions. However, an improvement is expected for the hm condition; although perhaps not as much as in aurora 2 as the mismatch is not as large.

Results are shown in Fig. 3. Contrary to expectations, there is a small improvement across the board, except for the Danish matched conditions. As expected, however, the improvement is most significant for the highest mismatch.

Aurora 4 is a noisy version of the well known wall street journal (WSJ) based SI-84 task. Aurora 4 goes back to using real noise artificially added to otherwise undistorted speech, but is large vocabulary (5000 words), hence covering the phone set thoroughly. As in aurora 2, both clean and multi-condition training sets are defined. However, rather than define tests at particular SNRs, 14 individual

enumerated tests are specified; these are summarised in Table 1.

Although a test harness was made available by Parihar et al. (2004), other authors have written their own (e.g., Au Yeung and Siu, 2004). In the present experiments, a scheme in the spirit of that of Parihar et al. (2004), but using HTK, was used.

To better reflect a typical WSJ system, the 16 kHz data were used with a 400 point DFT and 40 bank mel filter. Other parameters were as in the 8 kHz experiments. Results are shown in Fig. 4 (Sennheiser microphone) and Fig. 5 (second microphone). A priori, from the aurora 2 result, one would not expect the multi-condition results to vary much between SNR and energy based PLPs. The added noise is in the range 5–15 dB, however, which is within the range in which SNR features have been shown to afford an improvement. In this sense, the clean training results should be better for SNR based PLPs.

In practice, the a-priori expectations are borne out quite well.

### 5.7. Rich text

The SNR-cepstrum was briefly evaluated in the context of meeting room recognition. The baseline was the AMIDA RT06 system of Hain et al. (2006). Only the first pass was evaluated, and only the IHM (individual headset microphone). At an early stage, it was clear that the results from the SNR-cepstrum were no better than those from the baseline, and further experiments were abandoned.

In fact, this result is broadly what would be expected a-priori given the aurora 2 results. The training and test condition are matched, and the SNR is quite high; perhaps better than the notional 15 dB threshold.

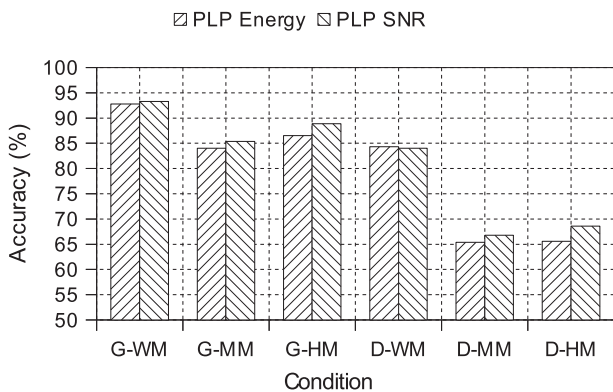


Fig. 3. PLP results on aurora 3 database. The G and D prefixes refer to German and Danish respectively.



Table 1  
Test set composition for aurora 4.

Microphone	Clean	Noise added between 5 dB and 15 dB					
		Car	Babble	Restaurant	Street	Airport	Train
Sennheiser	1	2	3	4	5	6	7
Second	8	9	10	11	12	13	14

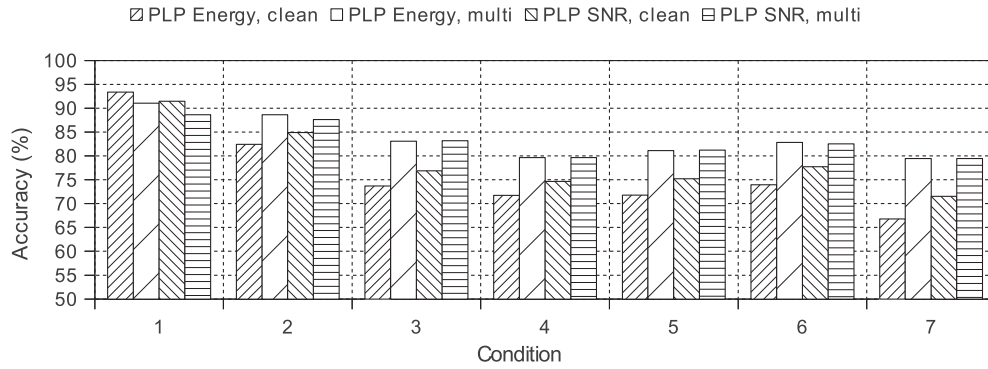


Fig. 4. PLP results on aurora 4 database – Sennheiser microphone.

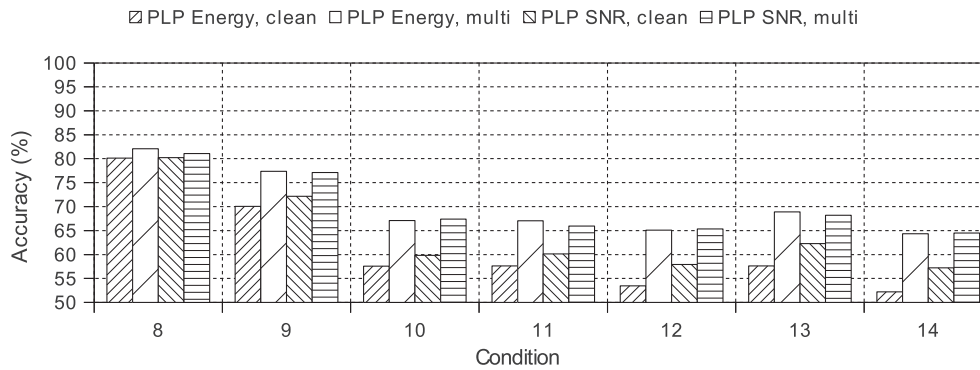


Fig. 5. PLP results on aurora 4 database – second microphone.

5.8. Experimental conclusions

The hypotheses are hence proven:

1. PLP features appear to benefit from SNR spectra in the same way as MFCC have been shown to do. At least on aurora 2, the results are better than for MFCCs.
2. Predictions made on the basis of aurora 2 results carry over to real noisy data, and to a large vocabulary system.

6. Discussion

6.1. State of the art

The experiments show that SNR-spectrum based features can be beneficial in noisy environments when there is a mismatch between the training and testing conditions. Garner (2009) also showed that such fea-

tures out-perform various types of spectral subtraction. No other comparison is made with other noise robustness techniques. Rather, the use of standard databases means the results can be readily compared with those in the literature. No claim is made that the SNR-spectrum gives state of the art results. For instance, Li et al. (2007) report considerably better results on aurora 2.

6.2. Analysis

That the SNR-spectrum performs well is a curious result since there is not a large theoretical difference between SNR spectrum features and energy spectrum features when CMN is used. The difference is that the SNR spectrum features normalise before the filter-bank, whereas CMN works after it.

If the filter-bank weights for a single bin are denoted by  $w_1, w_2, \dots$ , the SNR features presented to the decoder are of the form

$$\log(1 + w_1 \xi_1 + w_2 \xi_2 + \dots), \quad (29)$$

whereas the energy based features are *closer* to the form

$$\log\left(1 + \frac{w_1(s_1 + n_1) + w_2(s_2 + n_2) + \dots}{w_1 n_1 + w_2 n_2 + \dots}\right). \quad (30)$$

In broadband noise,  $\forall f: s_f \ll n_f$ , both expressions clearly reduce to the same value ( $\log 1$ ). However, if the noise is isolated to a particular bin,  $f$ , then only one term in the first expression will approach zero. In the second, the whole expression will reduce. It follows that the noises in the experimental conditions are suitably coloured for this effect to be significant.

These results are complementary to those of Lobdell et al. (2008), who also find advantages associated with AI features, albeit working after the filter-bank, and without cepstral normalisation.

### 6.3. PLP power law

One corollary of the aurora 2 experiments is that the cube root compression of Stevens (1957) normally used in PLP is not beneficial in the presence of noise. Whilst it is not the object of this study to investigate optimal PLP parameters, one hypothesis is as follows:

The compression affects the relative contribution of large and small spectral values in the LP calculation. Higher powers favour the higher values. The smaller power of 0.33 in PLP will enhance the contribution of smaller spectral values. The smaller values are likely to be noise. It follows that compression is in general not a noise robust operation. This issue is related to the SNR spectrum in that the SNR calculation can reduce noise peaks.

It can be tentatively concluded that additive noise is a more dominant concern than optimal compression in the present experimental conditions.

## 7. Conclusions

SNR-spectrum features for ASR have several practical and mathematical advantages over the more usual spectral power features. The naive SNR estimate is actually the optimal estimate under a fairly rigorous Bayesian analysis, and the framework leaves room for further incorporation of prior information, as is common recently in ASR. SNR features combined with CMN and CVN perform well in noisy conditions, especially when the SNR is below 15 dB.

The SNR-spectrum combined with the usual cepstral processing can be seen as an independent derivation of the articulation index. This also leads to insights into how to handle the noise tracker. Certainly the empirically optimal configuration is one with no hyper-parameters. The SNR-spectrum is also closely related to features known to be beneficial in speech enhancement.

Experiments on artificial and restricted data give results that appear to generalise to real and less restricted data.

Whilst no effort has been made to approach state of the art noise robustness figures, the SNR-spectrum appears complementary to techniques producing such results.

## Acknowledgements

This work was supported by the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

The author is extremely grateful to the four anonymous reviewers for their time and comments during the review process, especially regarding the presentation of results.

## References

- Acero, A., 1990. Acoustical and environmental robustness in automatic speech recognition. Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.
- Acero, A., Stern, R.M., 1990. Environmental robustness in automatic speech recognition. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, vol. 2, pp. 849–852.
- Allen, J.B., 1994. How do humans process and recognize speech? IEEE Trans. Speech Audio Process. 2 (4), 567–577.
- Allen, J.B., 2005. Consonant recognition and the articulation index. J. Acoust. Soc. Amer. 117 (4), 2212–2223.
- Au Yeung, S.-K., Siu, M.-H., 2004. Improved performance of Aurora 4 using HTK and unsupervised MLLR adaptation. In: Proc. Internat. Conf. on Spoken Language Processing, Jeju, Korea.
- Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27, 113–120.
- Cohen, I., 2005. Relaxed statistical model for speech enhancement and a priori SNR estimation. IEEE Trans. Speech Audio Process. 13 (5), 870–881.
- de la Torre, A., Peinado, A.M., Segura, J.C., Pérez-Córdoba, J.L., Benítez, C., Rubio, A.J., 2005. Histogram equalization of speech representation for robust speech recognition. IEEE Trans. Speech Audio Process. 13 (3), 355–366.
- Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6), 1109–1121.
- ETSI, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms, ETSI Standard 202 050, ETSI, v1.1.1, 2002.
- Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. 29, 254–272.
- Garner, P.N., 2009. SNR Features for automatic speech recognition, In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Merano, Italy.
- Garner, P.N., Dines, J., 2010. Tracter: a lightweight dataflow framework. In: Proc. Interspeech, Makuhari, Japan.
- Hain, T., Burget, L., Dines, J., Garau, G., Karafiat, M., Lincoln, M., Vepa, J., Wan, V., 2006. The AMI meeting transcription system: progress and performance. In: Proc. NIST RT06 Spring Workshop.
- Hain, T., Burget, L., Dines, J., Garner, P.N., El Hannani, A., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., 2010. The AMIDA 2009 meeting transcription system. In: Proc. Interspeech, Makuhari, Japan.
- Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.

- Hermansky, H., Morgan, N., 1994. RASTA processing of speech. *IEEE Trans. Speech Audio Process.* 2 (4), 578–589.
- Hirsch, H.-G., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: *ISCA ITRW ASR2000 “Automatic Speech Recognition: Challenges for the Next Millennium”*, Paris, France.
- Lathoud, G., Magimai-Doss, M., Mesot, B., Bourlard, H., 2005. Unsupervised spectral subtraction for Noise-Robust ASR. In: *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005.
- Lathoud, G., Magimai-Doss, M., Bourlard, H., 2006. Channel normalization for unsupervised spectral subtraction, *IDIAP-RR 06-09*, Idiap Research Institute, URL <<http://publications.idiap.ch>>.
- Li, J., Deng, L., Yu, D., Gong, Y., Acero, A., 2007. High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series. *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*. IEEE, Kyoto, Japan.
- Lindberg, B., 2001. Danish SpeechDat-Car Digits Database for ETSI STQ Aurora Advanced DSR, Technical Report, CPK, Aalborg University, URL <[http://aurora.hsnr.de/download/sdc\\_danish\\_report.pdf](http://aurora.hsnr.de/download/sdc_danish_report.pdf)>.
- Lobdell, B.E., Hasegawa-Johnson, M.A., Allen, J.B., 2008. Human speech perception and feature extraction. In: *Proc. Interspeech*, Brisbane, Australia.
- Makhoul, J., 1975. Linear prediction: a tutorial review. *Proc. IEEE* 63 (4), 561–580.
- Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.* 9 (5), 504–512.
- McAulay, R.J., Malpass, M.L., 1980. Speech enhancement using a soft decision noise suppression filter. *IEEE Trans. Acoust. Speech Signal Process.* 28 (2), 137–145.
- Moreno, P.J., 1996. Speech recognition in noisy environments. Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.
- Moreno, P.J., Raj, B., Stern, R.M., 1996. A vector Taylor series approach for environment-independent speech recognition. In: *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, Atlanta, US, pp. 733–736.
- Netsch, L., 2001. Description and Baseline Results for the Subset of the Speechdat-Car German Database used for ETSI STQ Aurora WI008 Advanced DSR Frontend Evaluation, STQ Aurora DSR Working Group input document AU/273/00, Texas Instruments, URL <[http://aurora.hsnr.de/download/sdc\\_german\\_report.pdf](http://aurora.hsnr.de/download/sdc_german_report.pdf)>.
- Parihar, N., Picone, J., Pearce, D., Hirsch, H.G., 2004. Performance Analysis of the Aurora Large Vocabulary Baseline System. In: *Proc. 12th European Signal Processing Conf.*, Vienna, Austria.
- C. Plapous, C. Marro, L. Mauuary, P. Scalart, A Two-Step Noise Reduction Technique, in: *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, Vol. I, Montreal, Canada, 289–292, 2004.
- Ris, C., Dupont, S., 2001. Assessing local noise level estimation methods: application to noise robust ASR. *Speech Commun.* 34 (1–2), 141–158.
- Segura, J.C., Benítez, M.C., de la Torre, A., Rubio, A.J., 2002. Feature extraction combining spectral noise reduction and cepstral histogram equalisation for robust ASR. In: *Proc. Internat. Conf. on Spoken Language Processing*, pp. 225–228.
- Stevens, S.S., 1957. On the psychophysical law. *Psychol. Rev.* 64 (3), 153–181.
- Van Compernelle, D., 1989. Noise adaptation in a hidden Markov model speech recognition system. *Comput. Speech Lang.* 3 (2), 151–167.
- Vikki, O., Laurila, K., 1997. Noise Robust HMM-Based Speech Recognition using Segmental Cepstral Feature Vector Normalization, in: *Robust Speech Recognition for Unknown Communication Channels*, ISCA, Pont-à-Mousson, France, 107–110.
- Viikki, O., Laurila, K., 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Commun.* 25, 133–147.