TKK Reports in Information and Computer Science
Espoo 2009 TKK-ICS-R19

STUDIES ON NOISE ROBUST AUTOMATIC SPEECH RECOGNITION

Kalle J. Palomäki, Ulpu Remes and Mikko Kurimo (Editors)

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

TKK Reports in Information and Computer Science
Espoo 2009 TKK-ICS-R19

STUDIES ON NOISE ROBUST AUTOMATIC SPEECH RECOGNITION

Kalle J. Palomäki, Ulpu Remes and Mikko Kurimo (Editors)

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O.Box 5400
FI-02015 TKK
FINLAND
URL: http://ics.tkk.fi
Tel. +358 9 451 1
Fax +358 9 451 3369
E-mail: series@ics.tkk.fi

ISBN 978-952-248-087-3 (Print)
ISBN 978-952-248-088-0 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)
URL: http://lib.tkk.fi/Reports/2009/isbn9789522480880.pdf

Multiprint Oy
Espoo 2009

Abstract: Noise in everyday acoustic environments such as cars, traffic environments, and cafeterias remains one of the main challenges in automatic speech recognition (ASR). As a reserch theme, it has received wide attention in conferences and scientific journals focused on speech technology. This article collection reviews both the classic and novel approaches suggested for noise robust ASR. The articles are literature reviews written for the spring 2009 seminar course on noise robust automatic speech recognition (course code T-61.6060) held at TKK.

Keywords: noise robustness, automatic speech recognition (ASR)

Acknowledgement: The work was supported by the Academy of Finland project "Auditory approaches to automatic speech recognition" and Adaptive Informatics Research Centre.

Preface

The research conducted on automatic speech recognition in the Adaptive Informatics Research Centre and the Department of Information and Computer Science in the Helsinki University of Technology has traditionally focused on large vocabulary continuous speech recognition (LVCSR) in morphologically rich and highly agglutinative languages such as Finnish, Estonian, and Turkish. Other issues like noise robustness began to receive more attention when the LVCSR system developed in the HUT speech group became accurate enough to be considered for practical applications. The first projects on noise robust speech recognition were launched in spring 2007. These projects as well as the new course on automatic speech recognition Mikko Kurimo lectured for the first time in autumn 2008 inspired the spring 2009 seminar course on noise robust automatic speech recognition.

Noise robustness has received much attention as a research theme in automatic speech recognition for in most realistic application environments, the speech signal is contaminated with background noise, reverberation, or both. The techniques proposed for improving noise robustness range from robust parametrisations to model-based compensation methods and microphone arrays. The methods operate on varied assumptions about the environment and noise type, and while some highly efficient methods like parallel model compensation or model-based feature enhancement methods are more sensitive to noise or background model estimation, some methods like robust feature extraction or missing data methods promise moderate improvements in all environments. Therefore, it is not sensible to compare or rank the methods in general, and even choosing a suitable method for a problem or an application requires recognising and understanding the basic assumptions and limitations in the method.

For the spring 2009 seminar on noise robust automatic speech recognition, we selected topics that would cover the different approaches and provide students with a complete view of the existing research work. In addition to the topics covered in the literature reviews collected here, the seminar hosted two guest lectures: Tuomas Virtanen from Tampere University of Technology presented his research on sound separation methods and Stevan Keraudy from Helsinki University of Technology presented his master's thesis topic histogram equalisation.

We would like to thank all the presenters and course participants for the interesting discussions we enjoyed during the seminar.

Otaniemi 22.9.2009

Kalle Palomäki, Ulpu Remes and Mikko Kurimo

CONTENTS

Microphone Arrays and Techniques in Automatic Speech Recognition

Sakari Tervo

Helsinki University of Technology
Department of Media Technology
email: sakari.tervo@hut.fi

Abstract

It is possible to achieve better speech recognition performance by using multiple microphones than with a single microphone. In this article, an overview of microphone array techniques for improving speech recognition systems is given. In addition, some typical microphone arrays that are used in speech recognition are shown. From the speech recognition point of view, traditional steered beamformers are reviewed as well as a recently developed pattern recognition approach. Moreover, an example of a multiple speakers speech enhancement algorithm is given.

1 INTRODUCTION

The main goal in using microphone arrays in automatic speech recognition systems is to achieve better speech recognition performance than with single microphone systems. During decades, several microphone techniques have been developed for this issue. Traditionally, steered beamformers are used to enhance the speech signals and then the automatic speech recognition is done to the enhanced signals [1]. Beamforming, originally developed for antenna arrays, is a class of methods where one aims to amplify signal or signals with more than one receiver. Basically, the steering can be done be physically steering the antenna, or by delaying the signals computationally. Also, adaptive versions of the steered beamformers have been developed. These methods use signal processing for speech enhancement without modifying the pattern recognition algorithms used in speech recognition. More recently, better results have been achieved by approaching the problem with pattern recognition approaches such as the likelihood maximizing beamforming methods, in which beamforming is adjusted based on top-down information from the recogniser [1]. In addition to these methods a multiple speaker scenario speech recognition system is introduced here.

In real situations there are several kinds of disturbance in the received speech signal. Firstly, there is additive noise that is considered to be caused by the recording system. Secondly, the enclosure or the environment causes distortion to the original signal. Finally, if there are more than one source in the enclosure, the sources that are not of interest are considered as disturbance signals.

The above explanation can be formulated as follows. In a room environment the sounds $s_k(t)$ traveling from the sound sources k to the receiver i are affected by the path $h_{i,k}(t)$ and the measurement noise $w_i(t)$:

$$p_i(t) = \sum_{k=1}^{K} [h_{i,k}(t) * s_k(t)] + w_i(t), \tag{1}$$

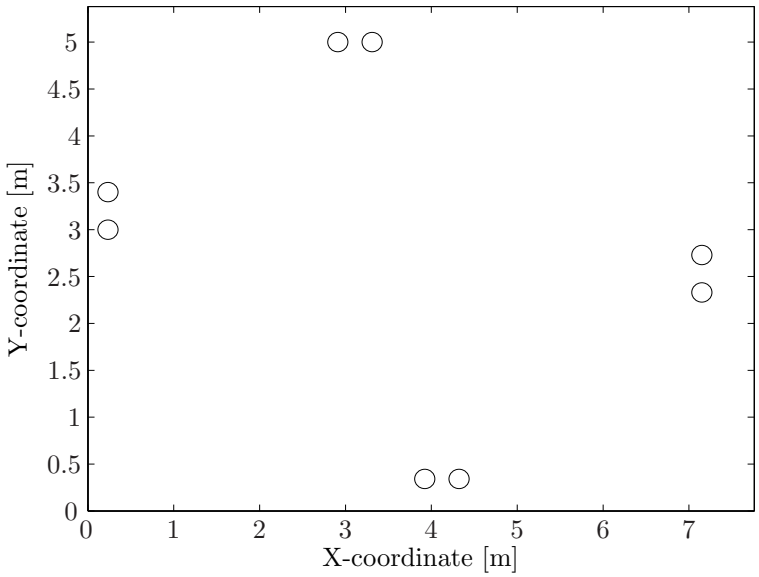

Figure 1: Typical small room environment recording setup.

where $*$ denotes convolution, p is sound pressure in the microphone, and K is the number of sources. Typically the complexity of the room environment is described by the well-known parameter T_{60}, i.e. the reverberation time, which can be estimated if $h_{i,k}(t)$ is known. The relation between $p_i(t) - w_i(t)$ and $w_i(t)$ is referred as signal-to-noise ratio.

Depending on the complexity of $p_i(t)$ one has to use different approaches to achieve robust speech recognition. For example if there is no reverberation, i.e. $h_{i,k}(t) = \delta(t - \tau)$ for all i, where τ is some constant and number of sources is one, $K = 1$, the task would be to find the most robust approach against additive noise. Ideally one would like to have a speech recognition system that is able to deal with all the difficult situations separately and at one time, i.e. when say $K > 1$, $T_{60} > 0.5$ s, or SNR < 5 dB. In this review the interest is more on the robustness against reverberation and other sources than on the robustness against noise.

2 MICROPHONE ARRAYS AND SETUPS FOR SPEECH PROCESSING

2.1 Placement of the microphones

The placement of the microphones plays an important role in array processing. The placement of the microphones depends mostly on the application. For example if the application is hands free communication with mobile devices the microphones cannot be placed very far apart from each other due to the size of the devices. On the other hand, if the environment is fixed to a, say, meeting room then the microphones can be positioned smartly.

In a small room environment a generally accepted rule of thumb is to use a single microphone array per wall. One example where the microphone array is a microphone pair is shown in Fig. 1. This kind of setups were used for example in the Classification of

Events, Activities, and Relationships (CLEAR) evaluation [2] throughout many years.

However, the most popular arrays for speech recognition are probably linear arrays. Linear arrays have been used in the recording of several speech corpuses [1]. The reason for the use of linear arrays is probably the highly directive directivity pattern. Thus, the side slopes of the directivity pattern are small. In the case of a single linear array one achieves more information on the direction of the sound source than on the distance.

2.2 B-format

B-format microphones are useful in mobile devices or devices that are not tied to any specific environment, i.e. computers. B-format microphones are introduced here due to their popularity in recent research. Compared to linear arrays they more compact and can therefore be installed to smaller devices. It is also very easy to cope with acoustic echo when using B-format signals, since the sound coming from certain direction can be easily cancelled.

First order B-format signals can be recorded with for example Soundfield microphones or with microphone pair measurements. The recording with Soundfield microphones is straightforward. Therefore, only the first order B-format signal recording with microphone pairs is formulated here. A square grid with four microphones is used. The microphones are located in the corners and the diagonal of the square is assumed length d. More precisely, in Cartesian coordinate system, microphones 1 and 2 are located along the x-axis at (x_1, y) and (x_2, y) with $x_1 - x_2 = d$ and microphones 3 and 4 along the y-axis at (x, y_1) and (x, y_2) with $y_1 - y_2 = d$. The corresponding polar coordinate presentation of the Cartesian coordinates is denoted with azimuth angle θ and radius r.

On a certain axis a, the sound intensity is given in the frequency domain as

$$I_a(\omega) = \Re\{P_{\mathrm{W}}^*(\omega)U_a(\omega)\}, \tag{2}$$

where $P(\omega)$ and $U_a(\omega)$ are the frequency presentations of the sound pressure and of the particle velocity with angular frequency ω. In addition, $\Re\{\cdot\}$ and is the real part of a complex number and * denotes the complex conjugate [3].

The average sound pressure between the microphones in square grid can be calculated as [3]:

$$P_{\mathrm{W}}(\omega) \approx \frac{1}{4}\sum_{i=1}^{4} P_i(\omega). \tag{3}$$

In addition, the particle velocity on the x-axis is estimated as

$$U_{\mathrm{x}}(\omega) \approx \frac{-j}{\omega\rho_0 d}[P_1(\omega) - P_2(\omega)], \tag{4}$$

where d is the distance between the two receivers and $\rho_0 = 1.2$ kg/m^3 is the median density of the air and j is the imaginary unit. Now, the sound intensity can be estimated with Eq. (2) and the approximations in Eqs. (3) and (4). For obtaining the y-component of the sound intensity, the microphones 1 and 2 are replaced in Eq. (4) with microphones 3 and 4.

Kallinger et al. showed that the square grid microphone setup suffers from a systematic bias caused by the approximation in Eq. (4). The unbiased direction estimate can be obtained by compensating the direction with the inverse function of:

$$\theta_{\text{biased}} = \frac{\sin(\omega \frac{d}{2c} \sin(\theta))}{\sin(\omega \frac{d}{2c} \cos(\theta))}, \tag{5}$$

where $c = 343$ m/s is the speed of sound. The inverse function does not have any closed form solution so it has to be approximated e.g. with linear interpolation [3].

3 MICROPHONE ARRAY TECHNIQUES FOR SPEECH RECOGNITION

In this section two microphone array techniques that aim to improve speech recognition are introduced. First one treats the automatic speech recognition as a signal processing task and second one as a pattern recognition task.

3.1 Beamforming Techniques

One of the most researched acoustical source localization approaches is steered beamforming (see e.g. [4] and references within). In the concept of steered beamformers (SBF) in speech recognition the speech recognition algorithms are usually applied to the enhanced output signal of the SBF. This improves the recognition performance [1].

The simplest implementation of the steered beamformer is the delay and sum beamformer, where the received signals are simply delayed with a steering delay and added [5]:

$$P(\mathbf{r}) = \sum_{i=1}^{M} a_i p_i(t - \tau_i(\mathbf{r})) \tag{6}$$

where τ_n is the steering delay for each microphone, M is the number of microphones $a_n = 1/M$ is a weighting factor, and $\mathbf{r}$ is the steered location. The SBF output, P is the sound pressure in location $\mathbf{r}$. The steering delays are calculated as a function of location:

$$\tau_i(\mathbf{r}) = \frac{|\mathbf{r} - \mathbf{r}_i|}{c}, \tag{7}$$

where $\mathbf{r}_i$ is the location for microphone i.

Steered beamformer can be presented in the frequency domain as a filter and sum beamformer [4]:

$$P(\mathbf{r}) = \int \left| \sum_{i=1}^{M} W_i(\omega) H_i(\mathbf{r}, \omega) X_i(\omega) \right|^2 d\omega, \tag{8}$$

where M is the number of microphones, $W(\omega)$ is frequency weighting function, and $X_i(\omega)$ is the i^{th} signal $x_i(t)$ in frequency domain. The weighting can for example be the so-called phase transform weighting [5]:

$$W_i(\omega) = |X_i(\omega)|^{-1}. \tag{9}$$

$H_i(\mathbf{r}, \omega)$, the beamformer filter for each signal at each location $\mathbf{r}$ in frequency domain is

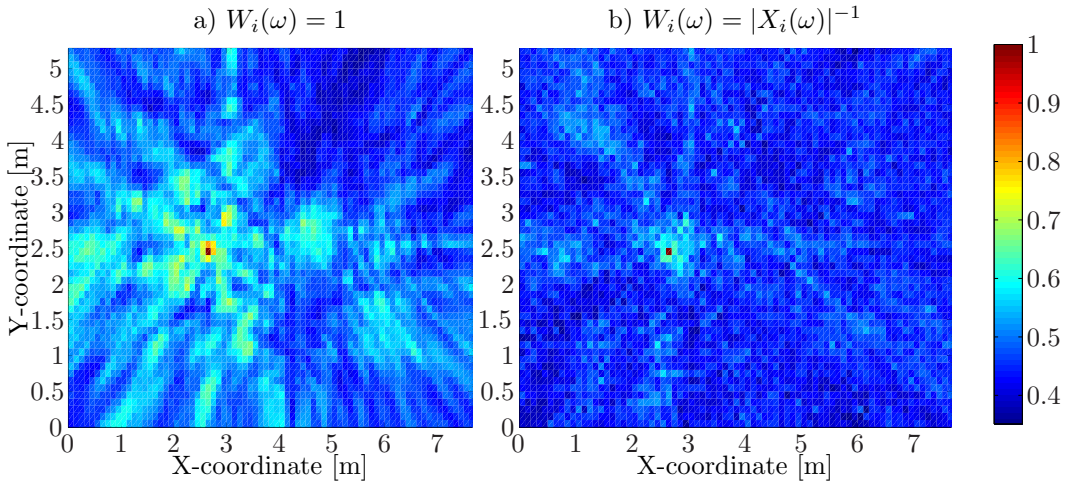

Figure 2: The steered beamformer output with two different filters. Red color indicates that there is more power or evidence in that area. The speaker is located at [2.4 2.6] m.

given as [4]:

$$H_i(\mathbf{r}, \omega) = \alpha_i e^{-j\omega c^{-1}(|\mathbf{r}-\mathbf{r}_i|-d_{\text{ref}})}, \tag{10}$$

where the gain for each received signal, may have uniform values $\alpha_i = 1/M$, and d_{ref} is chosen to be the center point of each microphone pair or array.

When SBF is used for single source location estimation the the maximum argument of the SBF function [5]:

$$\hat{\mathbf{r}}_s = \arg\max_{\mathbf{r}} P(\mathbf{r}), \tag{11}$$

is the location estimate. In Fig. 2 an example of the output for the steered beamformer with no filter (a) and phase transform filter is shown (b). The microphone setup was same as in Fig. 1. The room environment is simulated with the image source method [6] with 0.2 seconds of reverberation time. In addition, the source signal was male speech, the signal-to-noise ratio was set to 25 dB, and the used frequency band was from 100 Hz to 5 kHz.

In addition to the SBF above, many proposals to do the beamforming as an adaptive process has been made [1]. However, the SBF based methods have not been used widely in the speaker recognition systems due to their bad performance [1].

3.2 Likelihood Maximizing Beamforming

Likelihood maximizing beamforming is a class of methods that aim to find an optimal set of some parameters that maximizes the correct hypothesis, i.e. the correct word string. Here the general idea of this pattern recognition approach is reviewed. This section is based on [1].

Denote the observed feature vectors with $Z = [\mathbf{z}_1, \mathbf{z}_2, ...\mathbf{z}_T]$, which is dependent on speech and is a function of array processing parameters ξ, i.e. $Z(\xi)$. The array processing parameters are actually the coefficients of a FIR filter that corresponds to the impulse responses h_i. Automatic speech recognition aims to find the most likely word string

$\mathbf{c} = [c_1, c_2, ..., c_T]$ that generates the observed feature vectors, i.e.

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} P(Z(\xi)|\mathbf{c})P(\mathbf{c}) \tag{12}$$

$P(Z(\xi)|\mathbf{c})$ is the acoustic likelihood and $P(\mathbf{c})$ is the language likelihood. The goal is to find the set of parameters ξ that maximize the likelihood.

If the correct word string $\mathbf{c}_c$ is known, the likelihood function can be maximized with respect to the parameters:

$$\hat{\xi} = \arg\max_{\xi} P(Z(\xi)|\mathbf{c}_c)P(\mathbf{c}_c) \tag{13}$$

from where $P(\mathbf{c}_c)$ can be omitted.

In a hidden Markov model based speech recognition system, after some assumptions and simplifications (see [1] for details) the finding of the optimal parameter set formulates as:

$$\hat{\xi} = \arg\max_{\xi, \mathbf{s}\in\mathbf{S}_c} \left\{ \sum_i \log(P(Z(\xi)|s_i)) + \sum_i \log(P(Z(s_i|s_{i-1}, \mathbf{c}_c)) \right\} \tag{14}$$

where, s_i is a state in a state sequence $\mathbf{s}$, and all of the possible state sequences are denoted with $\mathbf{S}_c$ in a hidden Markov model for the sequence of words in the transcription $\mathbf{c}_c$. Optimization of the state sequence $\sum_i \log(P(Z(s_i|s_{i-1}))$ and the array parameters $\sum_i \log(P(Z(s_i|s_{i-1}, \mathbf{c}_c))$ is not considered here more. In [1] one solution for finding the state sequence and the array parameters is given. However, in this kind of non-linear optimization problem any global optimization tool can be used.

The likelihood maximizing beamforming achieves better performance than the traditional delay and sum beamformer, introduced in the previous section [1]. Seltzer reports that with unsupervised likelihood maximizing beamforming can achieve 27 % word error rate, when the delay and sum beamformer achieves 38.7 % word error rate with CMU-8 corpus.

4 BLIND SOURCE SEPARATION

Blind source separation is one type of speech enhancement process where two or more sources are separated from each other using information that is often achieved from spatial or time and frequency domain. Next a blind source separation system is reviewed [7].

4.1 Blind Source Separation Using Convolutive Mixtures

After calculating the sound intensity vectors the source separation process proceeds as follows. The direction of the sound intensity is estimated:

$$\theta(\omega) = \arctan 2\left\{ \frac{I_y(\omega)}{I_x(\omega)} \right\} \tag{15}$$

where arctan 2 is the four quadrant arctangent, and $I_y(\omega)$, $I_x(\omega)$ are the x and y components of the sound intensity vectors.

Next, the most likely source positions are searched using the histogram of $\theta(\omega)$ and a mixture model which is formulated as the sum of distributions f:

$$p(\theta|\mu,\kappa) = a_o/2\pi + \sum_{\mathrm{m}=1}^{\mathrm{M}} a_{\mathrm{m}} f(\theta|\mu_{\mathrm{m}},\kappa_{\mathrm{m}}), \tag{16}$$

where, m indicates the index of a mixture, a_{m} is assumed to be constant for all m, and $a_0/2\pi$ corresponds to the noise floor. The distributions in [7] were of von Mises probability density functions:

$$f(\theta|\mu_{\mathrm{m}},\kappa_{\mathrm{m}}) = \frac{e^{\kappa_{\mathrm{m}}\cos(x-\mu_{\mathrm{m}})}}{2\pi\mathcal{I}(0,\kappa_{\mathrm{m}})}, \tag{17}$$

where κ_{m} is the concentration, μ_{m} is the mean, and $\mathcal{I}(0,\kappa_{\mathrm{m}})$ is the modified Bessel function of order 0. The distributions can be fitted to the histogram of $\theta(\omega)$ with any global optimization methods. In [7] the fitting was done with maximum likelihood estimation by giving discrete values to κ and μ. It should be noted that this algorithm does not estimate the number of sources.

The enhanced signal is then formed by beamforming to the maximum likelihood solutions μ_{m} and κ_{m}. The beamforming function is defined as:

$$\hat{s}_{\mathrm{m}}(\omega,t) = P_{\mathrm{W}}(\omega,t) J_{\mathrm{m}}(\theta(\omega,t);\omega,t), \tag{18}$$

where t is index for frame and

$$J_{\mathrm{m}}(\theta;\omega,t) = a_{\mathrm{m}} f(\theta|\mu_{\mathrm{m}},\kappa_{\mathrm{m}}) \tag{19}$$

is the directivity function.

Gunel et al report that with their method it is possible to achieve good separation in small rooms [7]. For example in a room with 0.32 seconds of reverberation time the method achieves about 21 dB signal-to-interference ratio, with 90 degree separation between three sources. Despite the good separation capabilities that these kind of blind source separation methods provide, they cannot provide good dereverberation [1]. Thus, some other methods have to be used in addition to get good speech recognition performance.

5 CONCLUSIONS

Three microphone array techniques for improving the speech recognition performance were reviewed. Traditional sum and delay, maximum likelihood beamforming as well as blind source separation algorithms were introduced. Some typical microphone array setups for speech recognition were shown.

6 REFERENCES

[1] M. Seltzer, "Microphone Array Processing for Robust Speech Recognition", Ph.D. thesis, Carnegie Mellon University (2003).

[2] R. Stiefelhagen, K. Bernardin, R. Bowers, R. Rose, M. Michel, and J. Garofolo, "The CLEAR 2007 Evaluation", Multimodal Technologies for Perception of Humans, Joint Proceedings of the CLEAR 1–34 (2007).

[3] M. Kallinger, F. Kuech, R. Schultz-Amling, G. del Galdo, J. Ahonen, and V. Pulkki, "Enhanced Direction Estimation Using Microphone Arrays for Directional Audio Coding", in *Proc. Hands-Free Speech Communication and Microphone Arrays*, 45–48 (2008).

[4] E. Lehmann, "Particle Filtering Methods for Acoustic Source Localisation and Tracking", Ph.D. thesis, The Australian National University (2004).

[5] J. DiBiase, H. Silverman, and M. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications*, chapter 8. Robust Localization in Reverberant Rooms (Springer) (2001).

[6] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics", J. Acoust. Soc. Am **65**, 943–950 (1979).

[7] B. Gunel, H. Hacihabiboglu, and A. Kondoz, "Acoustic Source Separation of Convolutive Mixtures Based on Intensity Vector Statistics", IEEE Transactions on Audio, Speech, and Language Processing **16**, 748–756 (2008).

A Review of Missing-Feature Methods in Automatic Speech Recognition

Based loosely on paper of B. Raj and M. Stern: *Missing-Feature Approaches in Speech Recognition*. IEEE Signal Processing Magazine, Vol. 22, pp. 101-116, 2005

Okko Räsänen
Helsinki University of Technology
Department of Signal Processing and Acoustics
email: Okko.Rasanen@tkk.fi

Abstract

Current state-of-the art speech recognizers are based on Hidden Markov Model (HMM) architectures. Although their performance is sufficient for many applications in clean signal conditions, the recognition accuracy deteriorates badly when a mismatch exists between the training data and the speech being recognized. In almost all real world applications this mismatch is notable and the major source of this is the environmental background noise. Therefore a large amount of effort is being channeled to the development of methods that can increase noise robustness of the HMM recognizers, and hundreds of algorithms have already been developed. The so-called missing-feature approaches form one family of such methods. Missing-feature methods attempt to recognize unreliable spectral components from the input in a process called spectrographic masking and then use the reliability information to reconstruct original speech related spectral vectors or to modify the classifier in order to deal with the noisy data. This type of processing can significantly enhance speech recognition accuracy in difficult signal conditions. This work reviews the most essential missing-feature approaches reported in automatic speech recognition literature.

1 INTRODUCTION

Automatic speech recognition (ASR) has been a central engineering research topic for decades. Despite the continuous technological advances, especially the emergence of HMM-based speech recognition systems and the countless number of improvements to the HMM-methodology (Gales & Young, 2008; Juang & Rabiner, 2005; Makhoul & Schwartz, 1994), the state-of-the-art speech recognition systems fall still far behind human performance in recognition capabilities (Lippman, 1997; Moore, 2005). This is especially true for noisy conditions, that is, conditions with both semi-stationary and transient background sounds mixed with the input speech that is being recognized. Humans can perform source separation with relative ease, and they can increase their

recognition performance by utilizing multimodal- and contextual cues to fill gaps in input stream with most suitable explanation for missing data when the linguistic content of the input itself is ambiguous. HMM-based recognizers, on the other hand, require a match between the features extracted from the training data (state models) and the features obtained from the input signals in the recognition. Because this ideal match is rarely a reasonable assumption in real world applications, different types of methods have been developed to deal with noise in speech recognition processing. This paper reviews one family of such noise robust speech recognition approaches, namely *missing-feature methods*, that attempt to deal with partially corrupted spectral input. Since the HMM recognizers are the current state-of-the art in speech recognition, the missing feature approaches in this review all rely on HMM-based recognition systems unless mentioned otherwise.

Missing-feature method framework

The fundamental common assumption underlying approaches that are referred to as missing-feature approaches is that the spectral properties of input to a speech recognition system are partially corrupted due to environmental factors, and this impairment affects unequally different spectral elements in both discrete time and discrete frequency domain. The missing feature methods attempt to determine whether these different spectral components of the input are reliable representations of the speech being recognized (called *spectrographic masking*; section 2), and then take this reliability into account either by attempting ignoring the unreliable parts or by attempting to fill in the missing information using optimal estimation (section 3).

The assumption of partial corruption of the signals in time- and spectral domains often holds in real-world situations. At any moment of time, the majority of energy of clean speech is often concentrated into a very limited bandwidth compared to the entire bandwidth of human auditory perception. Moreover, the bandwidth required for intelligible speech communication is much smaller than the entire bandwidth of human hearing. Therefore the relevance of different spectral components of speech during different phonetic articulations differs greatly. On the other hand, loud environmental noise sources are often distributed differently than speech over auditory frequency range and most broadband sounds are temporally very limited in nature (e.g., loud impulse sounds) or low in intensity. When this noise is added to the speech signal, it is likely that the high-energetic spectral components of speech are relatively unaffected by lower intensity ambient noise, whereas the noise has greater impact on the parts of the spectrum that are low in speech intensity, causing unwanted spectral components to appear in feature vectors used for recognition (e.g., FFT or Mel

representations[1]). Due to redundancy of normal speech spectrum, it is also possible to some degree to obtain knowledge about high-energetic spectral structures (e.g., formants) even if their frequency bands are entirely masked with interfering high-intensity noise. Band-specific processing and reconstruction of the input are also related to human listeners, since it is well known that humans process auditory input at several parallel bands (Moore, 1997), and this band separation plays important role in both segregation of different audio streams (Bregman, 1990) and, e.g., learning of novel sound patterns (Newport & Aslin, 2004).

We will start with a description of how reliability of input data components can be estimated. Then more detailed descriptions of the two main missing-feature approaches in recognition, *feature-vector imputation* and *classifier modification*, will follow. A number of additional approaches are also reviewed that differ from the needs of mainstream continuous density HMM recognition. The description of the methodology will be strongly based on the work of Raj & Stern (2005) and follow the lines of their research, although some methods are described in more detail while others receive less attention.

2 ESTIMATING THE RELIABILITY: SPECTROGRAPHIC MASKS

First and possibly the most difficult aspect of the missing feature methods is the identification of the reliable and unreliable spectral components. The process of labeling spectral components in signal magnitude representation, or spectrogram, $Y(\omega,t)$ in terms of their reliability is called spectrographic masking, since in most simple case it divides the spectrogram of speech into a reliable and an unreliable parts. Naturally, this binary decision ("*hard masking*") does not have to be the case and scalar reliability measures ("*soft masking*") can be given to each spectral component in more sophisticated techniques if the next levels of processing can utilize this information efficiently in their decisions. A theoretical upper quality limit, and therefore a reference in evaluation of the proposed methods, is often considered to be the exact value of SNR in each component, which can be then used to construct an *oracle mask* (Cerisara et al., 2007). Naturally, this knowledge about actual SNR in each component is not available to the system in real world applications and the goal of a masking algorithm is to estimate prevailing signal conditions.

Spectrographic masking is a very important step in recognition process since it determines entirely how spectral frames are processed after mask estimation in missing data methods. Different approaches to masking can be divided to a handful of classes depending on the used criteria. Raj & Stern (2005) divide spectrographic masking

[1] Note that dealing with band specific corruption does not work directly for cepstral analysis. See, e.g., van Hamme (2004) for missing data techniques in cepstral domain.

methodology to SNR-based estimation and to methods that use some other criteria than SNR. Another division, proposed by Cerisara et al. (2007), is to have methods that mimic, or are inspired by the processing in human ear, and methods that are based on signal processing and stochastic modeling. Several major approaches to spectrographic masking will be discussed in the following sub-sections. In addition to the reviewed techniques, a number of other approaches have been proposed, including the use of neural networks (Potamitis et al., 2000), independent component analysis (Potamitis et al., 2001), and feature vector segmentation based on topological relationships in the spectrum for a set of features (Bach & Jordan, 2005), but are not included in this review due to lack of space.

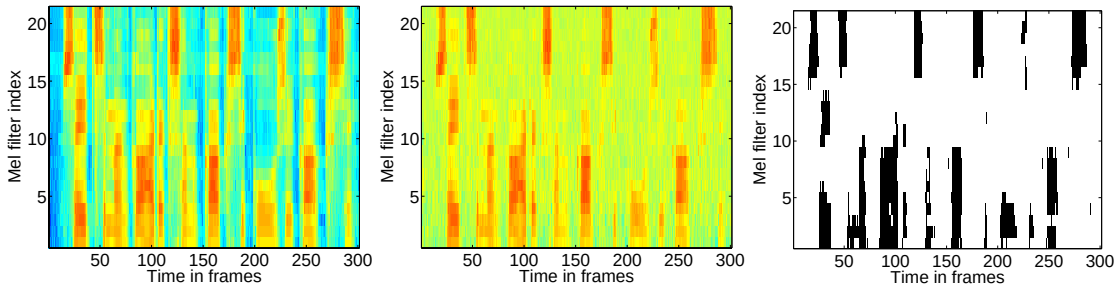

Fig. 1: A clean speech sample (a) is corrupted by pink noise to an SNR of 10dB (b). An SNR-based spectrographic mask (c).

2.1 SNR based estimation of spectrographic masks

The SNR based estimation of reliability of spectral components requires knowledge of energy of both speech signal and the corrupting noise. One possibility is to estimate the background noise level from sections of input that do not contain speech, non-speech sections determined by, e.g., voice activity detection or by assuming that first frames of any utterance are silence (Vizinho et al., 1999). However, it is shown that voice activity detection is often difficult to tune correctly for varying signal conditions and becomes unreliable at low SNR levels (Meyer et al., 1997; Sohn et al., 1999; McKinley & Whipple, 1997; see Cohen, 2003). These methods also rely on the assumption that the background noise signal is relatively stationary during the processed utterance so that the estimates of noise made during non-speech are valid for entire length of speech too.

To account for possible small changes in noise level during speech, a simple recursion can be utilized that initializes noise spectrum to the values obtained from first few frames of the utterance and then updates the band specific values if the change in energy between frames is sufficiently slow (fast changes are assumed to occur due to onsets of speech; Raj & Stern, 2005):

$$\hat{N}_p(m,k) = \begin{cases} (1-\lambda)\hat{N}_p(m-1,k) + \lambda Y_p(m,k) & \text{if } Y_p(m,k) < \beta\hat{N}_p(m-1,k) \\ \hat{N}_p(m-1,k) & \text{otherwise} \end{cases} \quad (1)$$

where $\hat{N}_p$ is the estimated noise power, Y_p is the signal power, m is the time frame, and k is the number of spectral coefficient. Typical values are $\lambda = 0.95$ and $\beta = 2$, where λ determines how quickly the noise variation is being tracked (Raj & Stern, 2005). Also improved (and more complicated) recursive averaging algorithms have been developed that utilize adaptive spectral and temporal smoothing, enabling soft transitions in noise level estimate between speech activity and non-activity. These have been shown to lead to improved performance in non-stationary and low SNR conditions (Cohen, 2003).

After the background noise level is determined, reliable and unreliable spectral components have to be determined according to some rule. A so-called negative energy criterion (El-Maliki & Drygajlo, 1999) assigns a spectral component as unreliable if its energy is less than the estimated noise energy in the corresponding frame:

$$\left|Y_p(m,k)\right| \le \left|\hat{N}_p(m,k)\right| \quad (2)$$

According to Raj & Stern (2005), best mask estimates are obtained if negative energy criterion is combined with a SNR criterion that assigns all components having SNR greater than 0 dB as reliable. SNR of each spectral component can be estimated by using spectral subtraction (Boll, 1979), where noise estimate is subtracted from the power spectrum of the noisy speech to obtain clean speech power estimate:

$$\hat{X}_p(m,k) = \begin{cases} Y_p(m,k) - \hat{N}_p(m,k) & \text{if } Y_p(m,k) - \hat{N}_p(m,k) > \gamma Y_p(m,k) \\ \gamma Y_p(m,k) & \text{otherwise} \end{cases} \quad (3)$$

Parameter γ has a small value and is used to prevent clean speech power from going to negative values.

In addition to these basic SNR-based approaches, measures of harmonical properties of (voiced) speech signals are often combined to the SNR analysis techniques. For example, measuring the similarity of shape of adjacent pitch periods and the use of Spectral Autocorrelation Peak Valley Ratio-Residual (SAPVR), which examines the autocorrelation of the FFT of LPC residual, have been utilized (Chandra & Yantorno, 2002). The assumption is that harmonic overtones of a sound in a signal can be considered as originating from the same source (de Cheiveigne et al., 1995), and

therefore these bands have lower probability to become masked due to their higher energy. When this type of harmonicity information is utilized, a final spectrographic mask is computed as a weighted sum of harmonicity mask and SNR mask.

2.2 Bayesian estimation of spectrographic masks

Instead of determining reliable components of speech by estimation of SNR level at each time frame for each separate frequency band, it is possible to use Bayesian decision-making approach to model clean speech and noisy speech features as probabilistic models and then match new input against these models.

One straightforward approach is to estimate probability density distributions for noise and clean speech in order to have a probability value for each spectral component that estimates whether noise energy level exceeds clean speech. This approach has clear improvement over normal spectral subtraction (e.g., 20 % recognition rate improvement in non-stationary helicopter noise; Renevey and Drygajlo, 1999; see also Renevey & Drygajlo, 2001). However, this approach relies on the assumption that the noise can be modeled with Gaussian or uniform distributions in beforehand.

Another approach has been introduced by Seltzer et al. (2004) that does not make any assumptions regarding the corrupting noise signal. It relies on classification based on a set of features that represent characteristic properties of speech. The feature set is divided to represent voiced and unvoiced speech separately since the production mechanisms are very different, and also because estimation of masks for unvoiced speech is more difficult due to relatively lower energies of the speech segments. Voiced speech features include 1) comb filter ratio that measures energy at the harmonics compared to the energy outside harmonics, 2) autocorrelation peak ratio describing the ratio of the largest and the second largest peak of the autocorrelation function, 3) subband energy to fullband energy ratio, 4) kurtosis of the spectrum, 5) variance of the spectral subbands (or *flatness*), 6) subband energy to subband noise floor ratio estimated from histogram of frame energies over entire utterance, and 7) the classical spectral subtraction based SNR estimate. Since unvoiced segments are non-periodic, they are described without the pitch dependent features 1) and 2).

The algorithm utilizes two separate two-class Bayesian classifiers, one for voiced and one for unvoiced speech, classes being reliable and unreliable speech represented by a mixture of Gaussians with a single full-covariance matrix tied across all densities in the mixture. Moreover, a separate classifier was constructed for each spectral subband. Prior probabilities for the two-classes can be estimated from training data as a fraction of frames that have SNR above and below a given threshold. It is noteworthy that this probabilistic framework enables the use of probabilistic spectrographic masks instead of

binary decisions, providing information which can be utilized later in HMM recognition processing. Seltzer et al. (2004) show that this methodology obtains clear improvement over SNR-based masking where binary masks are mapped to probability values using a sigmoid function of SNR (see Barker et al., 2001, for sigmoid mapping).

2.3 Computational auditory scene analysis-based techniques

It is well known that humans can utilize differences in auditory information arriving to two different channels (left and right ear) in order to enhance sound source segregation and to spatially localize sound sources in their surroundings (also called auditory scene analysis, or ASA; see, e.g., Bregman, 1990). Two main components in sound localization are interaural time difference (ITD) and interaural level difference (ILD). They are partially complementary, ITD being more suitable for localization at low sound frequencies whereas ILD works better on high frequencies due to masking properties of human head and phase locking accuracy of nerve cells in the cochlea (Moore, 1997). However, also other grouping cues exist that humans are able to exploit in auditory stream segregation. These cues include *simultaneous fusion* and *sequential organization* (Bregman, 1990) that come close to original Gestalt psychology. The first one relates to harmonical relations of a sound, periodicity, and especially the simultaneous onset of the stimuli. The latter includes rules how sequentially perceived sounds are grouped depending on, e.g., their frequency structure and interstimulus delay. These cues are also available in single channel signals, a situation normally faced in ASR (Cerisara et al., 2007).

Computational auditory scene analysis (CASA) literally attempts to simulate ASA with computational modeling. It is very closely related to missing data techniques in speech recognition, since human like auditory scene analysis and sound source segregation, if successful, would naturally solve the mask estimation problem since sound components originating noise sources and speech sources would become determined in the process. Therefore CASA models are of great interest in this context.

Most monaural CASA algorithms exploit harmonicity of the speech signals as the main cue in sound source segregation, a topic that was already discussed briefly in section 2.1, and more examples of methods utilizing this type of information in relation to CASA can be found in Brown & Wang (2005). Some algorithms have also utilized other cues like frequency modulation (Mellinger, 1991) and common onset of sounds (e.g., Brown, 1992; Denbigh & Zhao, 1992), although harmonicity is still often strongly embedded also in these approaches.

For multi-channel CASA processing, Palomäki et al. (2004; see also Palomäki et al., 2001) have proposed a perceptually motivated model in the framework of missing feature recognition using two sound channels, especially designed for reverberant

listening conditions. The model, briefly introduced here, consists of two parallel pathways: a monaural and binaural pathway (fig. 2). The monaural pathway is dedicated for peripheral auditory processing that provides feature vectors for a conventional missing data speech recognizer. It performs frequency band separation to 32 bands using ERB scale for determining the center frequencies of a gammatone filterbank. Outputs of the filters are half-wave rectified and Hilbert envelopes are computed, smoothed, compressed (power of 0.3), and sampled every 10 ms to have channel energies for feature vectors.

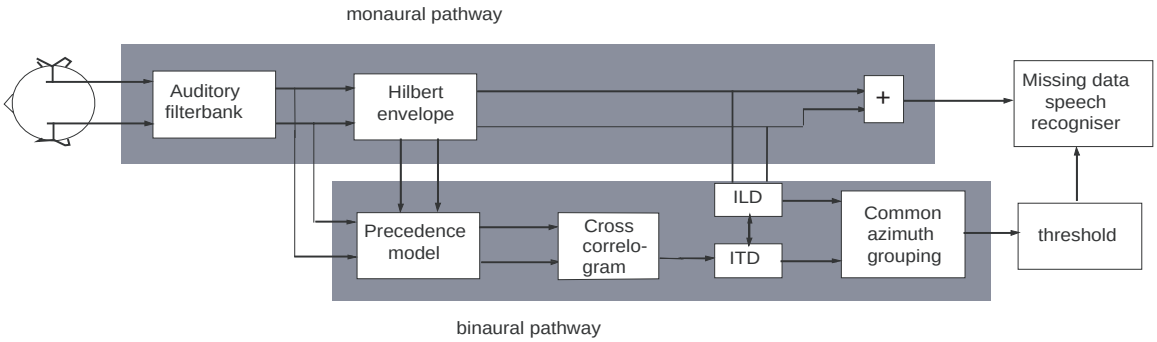

Fig. 2: Schematic diagram of the computational ASA model by Palomäki et al. (2004).

The binaural pathway is responsible for sound localization. Experimental data suggests that in reverberating conditions humans localize sound source according to the direction of the first wave front arriving to the ears, while reverberations only affect qualitative properties of the sound. This is called *precedence effect* (e.g., Blauert, 1997). In order to emphasize the onsets of arriving sounds and to de-emphasize later arriving echoes, Palomäki et al.'s model computes ITD from binaural signals that are first inhibited after sound onset using low-pass filtered Hilbert envelopes computed in the monaural channel (*precedence model* block in fig. 2). The ITD computation is performed using cross-correlation between the two channels arriving from the precedence model, leading to correlation values as a function of delay and frequency band. From delays, a straightforward warping to azimuth angle can be performed. Interaural level difference is computed separately for each frequency band starting from frequency of 2800 Hz.

In order to form a spectrographic mask, the azimuth information from ITD is pooled across channels, resulting in peaks corresponding to separate sound sources at different angles. An assumption is made regarding which of the peaks corresponds to speech, and information originating from corresponding angle is considered as reliable in the mask. The high frequency part of mask is then determined using ILD cue. Regions with ILDs that are consistent with ITD based azimuth estimate at specific

time instants are marked as reliable (one mask value). Additional energy threshold is included that sets reliability of a channel to zero if the ratio of instantaneous energy to the mean energy of the channel in a time window of 200 ms is lower than a predefined threshold. This improves mask quality in low SNR conditions.

When evaluated with TIDIGITS corpus (Leonard, 1984) using a model of small room acoustics (Allen & Berkley, 1979) and head-related impulse response to simulate spatial sound properties, significant improvements in speech recognition accuracy were obtained over normal MFCC features in different noise conditions including reverberation and interfering speech, interfering speech only, and rock music. For example, for interfering male speech in reverberant conditions (reverberation time $T_{60} =$ 0.3 s) and SNR of 10 dB, the word recognition rate with binaural mask was 83.4 % whereas only 47.6 % was obtained without noise compensation. Use of an oracle mask resulted in recognition rate of 94.6 %.

However, later research has pointed out that there is evidence regarding human hearing that grouping based on consistent ITD across frequency is weak (Edmonds & Culling, 2005), which is in contrast to use of ITD as an across frequency cue in this model.

2.4 Conclusions for mask estimation

Three different approaches for mask estimation were reviewed in this section. The estimation of spectral component reliability using knowledge of signal-to-noise ratio is effective if the SNR is known a priori (oracle masks) or (in practical applications) estimated reliably. However, good estimates for channel specific SNRs are extremely difficult to obtain in presence of non-stationary noise sources, since a priori assumptions about noise structure cannot be made. Bayesian decision theory can overcome noise-modeling problem by modeling the characteristics of the speech signal itself, combining information from spectral and temporal neighborhood to give a probabilistic estimate for reliability of each spectral component. Since this approach produces an actual probability value for each element (in terms of Bayesian thinking) instead of a binary classification, it can further enhance recognition performance in recognizers that are able to extend their search to this probabilistic space. Finally, it was shown that a psychoacoustically motivated approach to computational auditory scene analysis utilizing binaural processing can be used for sound source segregation, leading to highly increased recognition rates in difficult noise conditions if multi-channel auditory input is available.

3 DEALING WITH THE MISSING FEATURES IN RECOGNITION

After spectrographic masks are estimated, the reliability information has to be somehow utilized in the recognition process. According to Raj and Stern (2005), there are two major classes of methods for dealing with the unreliable data: *feature-vector imputation* methods that estimate most likely speech-related feature vectors from reliable spectral components, and *classifier modification* methods that modify the recognizer to deal with partially corrupted data. Both of these approaches will be discussed separately in the following two subsections.

3.1 Feature-vector imputation with cluster-based reconstruction

The aim of feature-vector imputation is to reconstruct original feature vectors representing the speech input from input vectors that contain spectral components corrupted by additive noise. An underlying assumption is that spectrographic masks identify the corrupted components and the remaining components can be treated as reliable descriptions of the speech signal.

One possibility for vector reconstruction is to use so-called *cluster-based reconstruction*. Spectrogram of the input signal power can be noted as $\mathbf{Y} = \mathbf{Y_u}+\mathbf{Y_r}$ where $\mathbf{Y_u}$ are the unreliable spectral components and $\mathbf{Y_r}$ are the reliable components, defined by the spectrographic mask. Now speech signal $\mathbf{X}$ that is being recognized can be divided to reliable and unreliable part $\mathbf{X} = \mathbf{X_u}+\mathbf{X_r}$, where $\mathbf{X_r} \approx \mathbf{Y_r}$ and $\mathbf{X_u}$ is unknown due to corrupting noise. However, since noisy (unreliable) components are assumed to have higher noise power than underlying speech power by definition, the unreliable components set an upper limit $\mathbf{X_u} \leq \mathbf{Y_u}$ to unknown speech components.

The distribution of clean speech vectors can be modeled with a Gaussian mixture

$$P(X) = \sum_v c_v (2\pi|\Theta_v|)^{-d/2} e^{-0.5(X-\mu_v)^T \Theta_v^{-1}(X-\mu_v)} \quad (4)$$

where d is dimension of the vector, c_v is the priori probability, μ_v is the mean vector, and Θ_v is the diagonal covariance matrix of the vth Gaussian. Now if a spectral vector $\mathbf{Y} = \mathbf{Y}_r+\mathbf{Y_u}$ is observed, the underlying speech component $\mathbf{X_u}$ can be estimated with bounded Maximum A-Posteriori (MAP) estimation:

$$\hat{X}_u = \arg_{X_u} \max\{P(X_u|X_r, X_u \leq Y_u)\} \quad (5)$$

For practical reasons, this is usually done by approximating the MAP estimate as a linear combination of conditional bounded MAP estimates

$$\hat{X}_u = \sum_v P(v|X_r, X_u \le Y_u) \arg_{X_u} \max\{P(X_u|X_r, X_u \le Y_u; \mu_v, \Theta_v)\} \quad (6)$$

where

$$P(v|X_r, X_u \le Y_u) = \frac{c_v P(X_r, X_u \le Y_u|v)}{\sum_j c_j P(X_r, X_u \le Y_u|j)} \quad (7)$$

Now obtained estimate for $\mathbf{X_u}$ and known $\mathbf{X_r}$ can be combined to a feature vector that is used normally in the further recognition processing (Raj & Stern, 2005).

Another possibility is to use covariance-based method for reconstructing the feature vectors. Expected values for speech vectors and corresponding covariance matrices can be trained using a clean speech corpus. Then all unreliable components in each spectral vector can be estimated with the bounded MAP procedure using the reliable components in the neighboring spectral vectors (see Raj & Stern, 2004, for a more detailed description of both approaches).

3.2 Classifier modification

Instead of recovering the structure of the feature vectors, it is possible to modify the classifier to deal with unreliable or incomplete data. This is called classifier-modification. Raj and Stern (2005) mention two popular approaches: *class-conditional imputation* (Josifovski et al., 1999) and *marginalization* (Cooke et al., 1999).

In clean speech conditions, HMM state emission distributions are of form $f(x|S)$. However, in noisy conditions, the feature vector x consists of reliable and unreliable components $x = x_r + x_u$ according to previous discussions. Since unreliable part is bounding the speech related values of x_u from above at x_{high}, the likelihood of data can be computed as

$$f(x_r|S) = \int_{x_{low}}^{x_{high}} f(x_r, x_u|S) dx_u \quad (8)$$

that is, by *marginalizing* the unknown elements, hence the name *bounded marginalization*. For continuous density HMM systems, state distributions can be modeled with Gaussians with diagonal covariance matrices:

$$f(x|S) = \sum_{k=1}^{M} P(k|S) f(x_r, x_u|k, S) \quad (9)$$

where $P(k|S)$ are the mixing coefficients, so the bounded marginalization can be formulated as:

$$f(x_r|S) = \sum_{k=1}^{M} P(k|S) f(x_r|k,S) \int_{x_{low}}^{x_{high}} f(x_u|k,S) dx_u \quad (10)$$

In *state based imputation*, also called *class-conditional imputation*, the missing components are estimated separately for each state, that is, the state specific density distributions are used to fill in the most likely missing information in case of that specific state and then these imputed state-specific feature vectors are used for normal likelihood computation. If densities of state distributions are assumed to be Gaussian like in (9), it can be shown that:

$$f(x_u|x_r,S) = \sum_{k=1}^{M} P(k|x_r,S) f(x_u|k,S) \quad (11)$$

Josifovski et al. (1999) measure recognition accuracy as a function of SNR with TIDIGITS continuous speech material. The effects of several different noise compensation were methods tested.

3.3 Other approaches

As an interesting another viewpoint to missing data recognition, the use of recurrent neural networks (RNN) has been studied in this framework and the first results seem promising. Parveen and Green (2001) modified a RNN architecture first introduced by Gingras and Bengio (1998) in order to enable it to perform simultaneous speech recognition and missing data imputation by having full recurrent links also from hidden layer to input layer (so called *Elman network*; see Elman, 1990). The feedback from hidden layer to the input-layer imputes unreliable spectral vector components based on missing components in previous input frames. Since imputation and recognition are embedded into the same process, this approach is a sort of hybrid of vector imputation and classifier modification.

Preliminary tests of Green and Parveen using this architecture with noisy TIDIGITS recognition indicate that the use of recurrent links for imputation of missing features improves recognition rates when compared to standard Elman RNN or imputation of features using unconditional mean values for missing components (Parveen & Green, 2001). Parveen and Green have also applied this similar type of Elman network for speech enhancement where noisy features are passed through the network to impute and/or correct corrupted data (Parveen & Green, 2004).

In addition to neural networks, another central issue is the noise compensation in systems with low computational capability where many sophisticated but computationally expensive methods are not possible (e.g., many hand-held devices). For this type of purpose, Renevey et al. (2001) have developed a weighting method that determines the reliability of each dimension in a feature vector in vector quantization. This reliability measure, or component weight, is utilized in distance computation when incoming feature vectors are compared to the centroids in VQ codebook. In their paper, they use normal Mahalanobis distance as the distance measure, but the weighting can be performed in principle to any other distance measure.

In their method, Renevey et al. set background noise threshold $\Theta(\omega)$ to the maximum magnitude value outside speech recognitions determined by voice activity detector:

$$\Theta(\omega) = \max_{t \notin speech} |Y(\omega,t)| \tag{12}$$

Using this threshold, they make normal division to reliable (signal energy higher than noise energy) and unreliable (signal energy lower than noise energy) subspaces. In addition, they make the following general assumptions: 1) *Speech and noise are additive in the magnitude spectral domain*. 2) *The highest the magnitude of the noisy signal, the smallest is the relative influence of the noise*. 3) *Under the assumption of additivity, features with a small magnitude are more representative of the speech signal than of the noise signal.* According to these, they formulate that in noisy speech subspace, the features with lowest magnitude are more representative of the clean speech signal than that with a highest magnitude, and in clean speech subspace features with the highest magnitude are more representative of the clean speech signal than features with value near threshold $\Theta(\omega)$. From these assumptions, a formula for weighting of feature vector components in noisy speech subspace can be written as:

$$\Psi_u(\omega,t) = 1 - \frac{|Y(\omega,t)|}{\max_{\omega \in unreliable}(|Y(\omega,t)|)} \tag{13}$$

and for the clean speech subspace:

$$\Psi_r(\omega,t) = \frac{|Y(\omega,t)|}{\max_{\omega \in reliable}(|Y(\omega,t)|)} \tag{14}$$

When tested with discrete density HMM and TIDIGITS corpus corrupted with noises from NOISEX noise database, the recognition rates were comparable, and sometimes

better, than results obtained standard codebook compensation methods using noise models and greatly improved from baseline quantization without noise compensation (Renevey et al., 2001).

3.4 Conclusions for dealing with missing data in recognition

The two predominant approaches for dealing with missing data in recognition were briefly introduced: feature-vector imputation and classifier modification. The former attempts to use reliable components of spectral vectors to estimate the remaining components corrupted by noise, whereas the latter either integrates over the possible values of missing component probability densities in the recognition phase or imputes the missing features separately for each state using the available state specific component distributions.

Not a single algorithm can be considered as a superior to others and some methods are more suitable for some specific type of noise situations and recognition tasks than some others. For example, experiments by Raj and Stern (2005) indicate that in the case of white noise, marginalization leads to significantly higher recognition rates than state-based imputation, and they both are significantly better than using only the basic spectral subtraction in the front end. On the other hand, spectral subtraction performs relatively well in car noise and GSM transmitted factory noise (Renevey, 2001). An interested reader is suggested to see Renevey (2001) for a comprehensive comparison of different missing-feature methods for several different noise conditions.

It was also pointed out that the computational complexity of continuous density HMM recognizers combined with computationally heavy mask estimation algorithms are not suitable for all ASR applications, and therefore computationally straightforward methods are needed and have been proposed.

4 SUMMARY

Missing-feature approaches in speech recognition attempt to improve speech recognition performance in noisy conditions, that is, in conditions that are nearly always faced in practical ASR applications. Higher noise robustness is achieved by first estimating reliability of each spectral component in the input frames in a process called spectrographic masking, and then utilizing this reliability information in order to fill in the missing spectral data or to modify the recognition process to be able to deal with the missing information.

Three major families of spectrographic masking were discussed. *Signal-to-noise ratio (SNR) based estimation* assumes a fixed or adaptive noise model for the noise corrupting the input signal and then uses this model to estimate SNR in each spectral

component. Components with insufficient SNR are considered as unreliable, or reliability is indicated as a scalar value mapped from the SNR itself. *Bayesian estimation* of spectrographic masks makes use of Bayesian probability theory. Classifiers are constructed and trained using clean speech and noisy speech in order to have probability distributions for each component. Additional features, e.g., ones that take into account the temporal neighborhood of the studied signal frame, can be included in the Bayesian decision process. Finally, *Computational Auditory Scene Analysis* (CASA) attempts to segregate sound sources using monaural and binaural grouping cues. Once a model for each sound source is obtained, the effects of noise on the studied speech signal are readily available.

Cluster-based reconstruction was introduced as a feature-vector imputation method, i.e., for filling in the corrupted components of each spectral vector. The approach uses either diagonal or full covariance Gaussian densities to estimate most likely values for unreliable components using the available reliable components and limits for the unreliable values. Feature vector imputation can be also done separately for each possible state in the recognizer using the probability densities tied to each state. This method is called *state-based imputation*.

Instead of attempting to recover corrupted features, c*lassifier modification approaches* change the recognition process itself according to the information provided in spectrographic masks. One possibility is to ignore the unreliable part by integrating over unreliable dimensions in probability distributions (*marginalization*) or to limit this integration to the values that are possible for the dimensions given the energy of unreliable noisy components (*bounded marginalization*).

All reviewed methods improve recognition accuracy in noisy conditions, although they may actually hinder recognition rates slightly on clean speech. Some methods (e.g., many SNR based methods) also make explicit assumptions regarding the properties of the noise, making them vulnerable to many non-stationary noise sources in real environments that are hard to model and predict, whereas others do not attempt to model noise itself at all and are more robust in varying conditions (e.g., Bayesian classification and CASA). Finally, it is important to note that some missing-feature approaches reviewed here are designed to work in frequency domain using logarithmic FFT densities, although the best speech recognition performances with clean speech have been obtained with continuous density HMMs using cepstral features that are incompatible with frequency band-specific processing. Therefore these methods are not directly applicable to recognizers that make the most of decorrelating properties of the cepstrum. However, this is not true for all methods and some, e.g., cluster based reconstruction methods (Raj et al., 2004), are capable to perform noise compensation also for cepstral features.

5 REFERENCES

J. Barker, M. Cooke, and P. Green: *Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise.* In Proc. Eurospeech'01, pp. 213-216, 2001

J. Blauert: *Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition).* MIT Press, Cambridge, MA, 1997

S. F. Boll: *Suppression of acoustic noise in speech using spectral subtraction.* IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 27, No. 2, pp. 113-120, 1979

A. S. Bregman: *Auditory scene analysis*. MIT Press, Cambridge, MA, 1990

G. J. Brown: *Computational auditory scene analysis: A representational approach.* Ph.D. Thesis, University of Sheffield, 1992

G. J. Brown and D. L. Wang: *Separation of speech by computational auditory scene analysis.* In J. Benesty, S. Makino, and J. Chen (Eds.) *Speech Enhancement,* Springer, New York, pp. 371-402, 2005

C. Cerisara, S. Demange, and J-P. Haton: *On noise masking for automatic missing data speech recognition: A survey and discussion.* Computer Speech and Language, Vol. 21, pp. 443-457, 2007

N. Chandra and R. E. Yantorno: *Usable speech detection using the modified spectral autocorrelationpeak to valley ratio using the LPC residual. In. Proc. IASTED International Conference on Signal and Image Processing, pp. 146-149, 2002*

I. Cohen: *Noise Spectrum Estimation in Adverse Environments: Improving Minima Controlled Recursive Averaging.* IEEE Transactions on speech and audio processing, Vol. 11, No. 5, pp. 466-475, 2003

M. Cooke, P. Green, L. Josifovski, and A. Vizinho: *Robust ASR with unreliable data and minimal assumptions.* In Proc. Robust'99, Tampere, Finland, pp. 195-198, 1999

A. de Cheiveigne, S. McAdams, J. Laroche, and M. Rosenberg: *Identification of concurrent harmonic and inharmonic vowels: a test of the theory of harmonic cancellation and enhancement.* Journal of Acoustical Society of America, Vol. 97, No. 6, pp. 3736-3748, 1995

P. N. Denbigh and J. Zhao: *Pitch extraction and separation of overlapping speech*. Speech Communication, Vol. 11, pp. 119-125, 1992

B. A. Edmonds and J. F. Cullings: *The spatial unmasking of speech: evidence for within- channel processing of interaural time delay.* Journal of Acoustical Society of America, Vol. 117, No. 5, pp. 3069-3078, 2005

J. L. Elman: *Finding structure in time.* Cognitive Science, Vol. 14, pp. 179-211, 1990

M. Gales and S. Young: The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, Vol. 1, No. 3, pp. 195-304, 2008

F. Gingras and Y. Bengio: *Handling Asynchronous or Missing Data with Recurrent Networks.* International Journal of Computational Intelligence and Organizations, Vol. 1, No. 3, pp. 154-163, 1998

L. Josifovski, M. Cooke, P. Green, and A. Vizinho: *State based imputation of missing data for robust speech recognition and speech enhancement.* In Proc. Eurospeech'99, Budabest, Hungary, pp. 2837-2840, 1999

B. H. Juang and L. R. Rabiner: *Automatic Speech Recognition – A Brief History of the Technology development*. Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005

R. G. Leonard: *A Database for Speaker-Independent Digit Recognition.* In Proc. ICASPP'84, Vol. 3, pp. 328-331, 1984

R. P. Lippmann: *Speech recognition by machines and humans*. Speech Communication, Vol. 22, pp. 1-15, 1997

J. Makhoul and R. Schwartz: *State of the Art in Continuous Speech Recognition*. In: Roe, D.B. & Wilpon, J.G. (Eds.), Voice Communication Between Humans and Machines, National Academy Press, Washington D.C., pp. 165-198, 1994

B. L. McKinley and G. H. Whipple: *Model based speech pause detection.* In Proc. 22th IEEE International Conference on Acoustics and Speech Signal Processing, ICASSP-97, Munich, Germany, pp. 1179-1182, 1997

D. Mellinger: *Event formation and separation in mmusical sound.* Ph.D. Thesis, Stanford University, 1991

J. Meyer, K. U. Simmer & K. D. Kammeyer: *Comparison of one- and two-channel noise-estimation techniques.* In Proc. 5th International Workshop on Acoustic Echo and Noise Control, IWAENC-97, London, pp. 137-145, 1997

C. J. Moore: *An Introduction to the Psychology of Hearing*. Fourth edition, Academic Press, California, 1997

R. K. Moore: *Spoken Language Processing: Piecing Together the Puzzle*. Speech Communication, Special Issue on *Bridging the Gap Between Human and Automatic Speech Processing*, Vol. 49, pp. 418-435, 2007

E. L. Newport and R. N. Aslin: *Learning at a distance I. Statistical learning of non-adjacent dependencies*. Cognitive Psychology, Vol. 48, pp. 127-162, 2004

K. J. Palomäki, G. J. Brown, and D. Wang.: *A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation.* Speech Communication, Vol. 43, pp. 361-378, 2004

K. J. Palomäki, G. J. Brown, and D. Wang.: *A binaural model for missing data speech recognition in noisy and reverberant conditions.* In Proc. Consistent & Reliable Acoustic Cues for sound analysis, CRAC01, Aalborg, Denmark, 2001

S. Parveen and P. D. Green: *Speech Recognition with Missing Data using Recurrent Neural Nets.* In Proc. Advances in Neural Information Processing Systems, NIPS*2001, 2001

S. Parveen and P. D. Green: *Speech Enhancement with Missing Data Techniques using Recurrent Neural Networks.* In Proc. ICASPP'04, pp. 733-736, 2004

I. Potamitis, N. Fakotakis, and G. Kokkinakis: *Impulsive noise removal using neural networks*. In Proc. ICASSP'00, Istanbul, Turkey, pp. 1871-1874, 2000

I. Potamitis, N. Fakotakis, and G. Kokkinakis: *Bayesian independent component analysis as applied to one-channel speech enhancement.* In Proc. International Conference on Artificial Neural Networks, Vol. 2130, pp. 593-600

B. Raj and M. Stern: *Missing-Feature Approaches in Speech Recognition*. IEEE Signal Processing Magazine, Vol. 22, No. 5, pp. 106-116, 2005

B. Raj and M. Stern: *Reconstruction of missing features for robust speech recognition*. Speech Communication, Vol. 43, No. 4, pp. 275-296, 2004

P. Renevey and A. Drygajlo: *Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition.* In Proc. Eurospeech'99, Budabest, Hungary, pp. 2627-2630, Vol. 6, 1999

P. Renevey and A. Drygajlo: *Detection of reliable features for speech recognition in noisy conditions using a statistical criterion.* In Proc. Consistent & Reliable Acoustic Cues for sound analysis, CRAC01, Aalborg, Denmark, 2001

P. Renevey, R. Vetter, and J. Kraus: *Robust Speech Recognition using Missing Feature Theory and Vector Quantization.* In. Proc. Eurospeech'01, pp. 1107-1110, 2001

P. Renevey: *Speech Recognition in Noisy Conditions Using Missing Feature Approach*", Ph.D. thesis, Lausanne, EPFL, 2000

M. L. Seltzer, B. Raj, and R. M. Stern: *A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition*. Speech Communication, Vol. 43, No. 4, pp. 379-393, 2004

J. Sohn, N. S. Kim, and W. Sung: *A statistical model-based voice activity detector.* IEEE Signal Processing Letters, Vol. 6, No. 1, pp. 1-3, 1999

H. van Hamme: *Robust speech recognition using cepstral domain missing data techniques and noisy masks.* In. Proc. ICASSP'04, Montreal, Canada, pp. 213-216, 2004

A. Vizinho, P. Green, M. Cooke, and L. Josifovski: *Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study.* In. Proc. Eurospeech'99, Budabest, Hungary, Vol. 5, pp. 2407-1410, 1999

Feature Extraction

Heikki Kallasjoki

Helsinki University of Technology
Department of Information and Computer Science
email: htkallas@cis.hut.fi

Abstract

Feature extraction is a crucial step in an automated speech recognition system, but conventional feature representations are not designed to be especially robust to noise. Better recognition rates, particularly in noisy conditions, can be achieved by using spectral estimation methods that are relatively invariant to noise. This article reviews alternative audio spectrum estimation methods based on linear predictive (LP) and minimum variance distortionless response (MVDR) modeling, in particular their suitability for speech recognition tasks. In addition, a different proposed feature representation based on time-domain filtering of the spectral frequency channels is presented.

1 INTRODUCTION

The extraction of an efficient set of features from a speech audio signal is an important step in an automated speech recognition system. The initial audio data contains a large amount of speaker- and situation-dependent variation that does not affect the meaning of what was said, and an ideal feature extraction system would discard as much as possible of this undesirable information. For example, a prominent feature present in a speech audio signal is the harmonic structure, which is mostly associated with the tone of voice used, and therefore is of no importance in a pure speech recognition scenario where only the textual content of speech needs to be recovered. Additionally, the selected features should ideally be relatively invariant to any noise present in the recording environment.

A very widely used feature set in speech recognition is the mel-frequency cepstral coefficient (MFCC) representation [1]. The MFCC representation is based on the cepstral coefficients of a short-term spectrum estimate of the speech signal, smoothed with a perceptually motivated filterbank. This perceptual smoothing does remove the speech harmonic structure to some extent, but the effect is limited. The resulting features are also not considered to be particularly robust in the presence of noise [2].

In this article, several methods of achieving more noise robust feature representations are reviewed. Generally the methods are variations of the MFCC method, utilizing more involved spectrum estimation than the simple periodogram used in the traditional MFCC. Section 2 gives a short overview of the commonly used MFCC features, as well as the major points of the linear predictive (LP) analysis, as background knowledge for the different methods described in section 3. Methods presented in 3.1 are based on LP modeling, while those in section 3.2 rely on the alternative minimum variance distortionless response (MVDR) modeling technique. Finally, section 3.3 describes the RASTA method, which involves time-domain filtering of the spectral channels.

2 BACKGROUND

2.1 Mel-frequency Cepstral Coefficients (MFCCs)

Mel-frequency cepstral coefficients are a very widely used feature representation for speech. An overview of a MFCC-based feature extraction algorithm is given below: [3]

1. An estimate for the power spectrum of the analyzed signal is computed, typically simply as the squared magnitude of the result of the FFT transformation.
2. The power spectrum estimate is smoothed with a perceptually motivated filterbank. In case of mel-cepstral analysis, the filters have a triangular shape, are spaced logarithmically, and applied to the logarithmic magnitude of the spectrum.
3. The discrete cosine transformation (DCT) is applied to the mel filterbank output, to get the cepstral coefficients. Typically only part of the cepstral coefficients are kept for analysis.

The logarithmic spacing of the filters in step 2 approximates the variable spectral resolution of the human hearing, while the logarithmic compression applied to the spectrum both resembles the intensity-loudness relationship of the human ear and reduces amplitude variations. The truncation of cepstral coefficients in step 3 further smoothes the spectrum, as it throws away higher-frequency components in the spectral domain.

2.2 Linear Prediction

Linear predictive analysis is a widely used tool for spectral modeling. In the linear prediction model, a sample x_n is estimated as: [4]

$$\hat{x}_n = -\sum_{i=1}^{p} a_i x_{n-i}, \tag{1}$$

where p is the model order, and $a_i \in \mathbb{R}$ are the linear prediction coefficients. The coefficient vector $\mathbf{a}$ is found by minimizing the cost function:

$$\mathcal{E}(\mathbf{a}) = \sum_{n=1}^{N+p} \varepsilon_n^2(\mathbf{a}), \quad \text{where } \varepsilon_n(\mathbf{a}) = x_n - \hat{x}_n \tag{2}$$

The plain LP spectrum is capable of removing some of the harmonic structure, but especially for middle- and high-pitched speech, it has a tendency of modeling the sharp contours of the harmonic structure as the model order is increased. The LP spectrum is also not especially robust in noisy environments. [5]

3 NOISE ROBUST FEATURE EXTRACTION METHODS

3.1 Linear Predictive Spectral Estimates

The plain LP estimate of the power spectrum models the original spectrum equally closely over all frequencies. However, the human auditory system is far from being so uniform. Notably, the spectral resolution is lower for high frequencies, and the sound power leading to equal perceived loudness is lower in the middle of the audible spectral range. Consequently, [6] proposes a perceptual linear predictive (PLP) approach to LP modeling in speech analysis.

In the PLP model, the signal autocorrelation estimates are computed from the output of a perceptually motivated filterbank, resembling the mel filters used in MFCC computation, although in addition to the change in spectral resolution, the PLP filterbank also emphasizes the spectrum samples using an equal-loudness curve representing the varying sensitivity of human hearing at different frequencies. A cubic-root compression step, analogous to the logarithmic scale used in the MFCC method, is also applied to the PLP filterbank output. After these steps, a conventional all-pole linear predictive model is constructed using the modified spectrum. In isolated word recognition experiments of [6], recognition accuracy rate of over 50 % was achieved with a PLP model of order 5, while the highest accuracy rate of any plain LP model never exceeded 40 %. Some results of PLP compared against the MVDR-based PMCC method are also given in section 3.2.

While the PLP features are somewhat more robust to noise, they are not especially designed for that purpose. There are also other approaches for improving the performance of linear prediction in speech recognition. The weighted linear prediction (WLP) method allows for providing a temporal weighting function which can be used to direct the LP model to emphasize particular temporal regions of the analyzed signal. Stabilized weighted linear prediction (SWLP) adds a transformation that guarantees the stability of the resulting all-pole model, and was recently used in the spectral estimation step of an MFCC-based system [7] as a candidate for noise robust speech features. The short-time energy function of the signal was used for the temporal weighting, as this causes the LP modeling step to focus on the high-energy sections of the speech signal that are relatively less affected by additive noise. The STE window also emphasizes the part of speech signal immediately after the glottal pulse, during which formants for voiced sounds are the most prominent.

3.2 Minimum Variance Distortionless Response (MVDR) Spectrum

Minimum Variance Distortionless Response (MVDR) modeling is an alternative method for estimating the upper spectral envelope, addressing several of the shortcomings of the conventional LP method. Notably, the MVDR spectral envelope estimate is much less prone to modeling the sharp contours of the harmonic structure of speech, even for higher-pitched speech where the harmonics are located more sparsely. Another useful property of the MVDR spectrum is that it can be efficiently computed from conventional LP coefficients. [5]

Conceptually, in the MVDR estimation method, the signal power at a given frequency ω_l is defined as the output power of a specific Mth order FIR filter $h_l(n)$. This filter is designed separately for the particular frequency ω_l and the particular input data being analyzed, to produce the least possible output power, with the constraint that it has a unit frequency response at frequency ω_l: [5]

$$H_l(e^{j\omega_l}) = \sum_{k=0}^{M} h_l(k) e^{-j\omega_l k} = 1 \tag{3}$$

However, it is not necessary to actually use a separate filter $h_l(n)$ for each frequency band. Using the Mth order LP model coefficients a_k and prediction error P_e, the Mth order MVDR spectrum $P_{\mathrm{MV}}^{(M)}$ can be written as: [5]

$$P_{\mathrm{MV}}^{(M)}(\omega) = \frac{1}{\sum_{k=-M}^{M} \mu_k e^{-j\omega k}}, \text{ where :} \tag{4}$$

$$\mu_k = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{M-k} (M+1-k-2i) a_i a_{i+k}^*, & \text{for } k = 0, \ldots, M, \\ \mu_{-k}^*, & \text{for } k = -M, \ldots, -1. \end{cases} \tag{5}$$

There are various ways the MVDR spectral envelope estimate can be used in feature extraction. [8] compares a straight-forward replacement of the periodogram in conventional MFCC computation with the MVDR estimate (MVDR-MFCC) against a more involved method, where the MVDR model coefficients are obtained from a perceptually smoothed spectrum (PMCC). [2] additionally presents experiments comparing the PMCC method with an alternative PMVDR method, where the perceptually scaled spectrum is computed with a warping function instead of a traditional mel filterbank.

The conventional LP coefficients used for deriving the MVDR coefficients are based on autocorrelation estimates, typically computed with the inverse FFT from the power spectrum estimate of the signal. In the PMCC method [8], the perceptual smoothing with the mel filterbank is done before the MVDR analysis, and the autocorrelation estimates are calculated from the smoothed spectrum, similarly to the PLP method. The smoothing leads to more reliable estimates, as the variance of the spectrum is diminished. The computational complexity of the method is also lower, due to the lower dimensionality of the mel filterbank output compared to the original FFT-based power spectrum estimate.

Results for a set of experiments, including both isolated word and continuous speech settings, comparing conventional MFCC with PLP, MVDR-MFCC and PMCC are presented in [8]. In an isolated digit recognition task using data recorded in a real car environment with a velocity of 60 mph, the achieved word error rates were 6.65 %, 6.22 %, 5.93 % and 4.92 % for MFCC, MVDR-MFCC, PLP and the PMCC methods, respectively. In the less noisy case of a velocity of 30 mph the results were similar, with smaller differences, while in the clean speech case recorded in a stationary car the differences were not significant. The continuous speech recognition experiments were limited to clean data, and did not include the MVDR-MFCC or PLP features, but the PMCC feature extraction achieved the best results, with an overall relative improvement of 12.8 % of the word error rate in tests using the Wall Street Journal data.

As an alternative to perceptual smoothing using mel-scaled filters, [2] presents the PMVDR method where the perceptually modified autocorrelation values are obtained from a directly warped FFT power spectrum, avoiding the loss of information inherent in the mel filterbank operation. Normally, in addition to the perceptual considerations, the filtering is also desirable for its ability to remove to some extent the prominence of the harmonic structure and other undesired speaker-dependent variation in the signal. Since the MVDR spectral estimate is capable of more accurately tracking the upper spectral envelope, and therefore less affected by the harmonic structure, the filtering step is not really necessary, and the MVDR modeling can be done on a directly warped spectrum. In a digit recognition task using the CU-Move database containing noisy speech from real car environments, overall word error rates of 11.12 %, 9.87 % and 7.74 % were achieved by the MFCC, PMCC and PMVDR methods, respectively. PMVDR, with a perceptual warping factor optimized specifically for the CU-Move task, achieved a lower word error rate of 7.11 %. Similar results were reached in another isolated word test using the SUSAS database of speech produced under stress.

3.3 RASTA Processing

The primary aim of the RASTA feature representation is to ignore slowly varying components of the speech signal directly in the spectral estimate. The concept is partially motivated by human auditory perception of speech, which seems to depend more on the spectral difference between consecutive sounds than on the absolute shape of the power spectrum itself. Another basis for the model is that in many situations the environmental background noise is constant or only slowly varying, and therefore a feature representation that discounts slowly varying factors will be more robust to it. [9]

In [9], the RASTA approach has been applied to perceptual linear prediction (PLP) based feature extraction scheme. The RASTA-specific processing starts with the PLP critical-band power spectrum, which resembles the mel filterbank output in conventional MFCC computation. The separate power spectrum samples are then transformed using a fixed compressing nonlinear transformation. After compression, each spectral channel is filtered in the time domain, together with corresponding power spectrum samples from preceding frames, using a suitable band-pass filter. The aim of the band-pass filter is twofold. The high-pass portion is expected to reduce the effect of convolutional noise, as it filters out the constant and slowly varying components on the particular frequency band. In contrast, the purpose of the low-pass filtering is to discard excessively fast changes between successive frames, under the assumption that they are merely artifacts caused by the spectral analysis. Finally, the filtered power spectrum is transformed using an expanding transformation matching the earlier compression stage, and the result is processed following the basic PLP method.

The RASTA processing can be compared to the common cepstral mean subtraction (CMS) operation often used in continuous speech recognition systems, since they both affect the spectral domain in roughly the same way. The primary difference between RASTA and CMS is that the typical length of a CMS window is long enough for the operation to only effectively remove a constant bias, while the filters used in RASTA

processing have a larger effect on the spectrum values. According to [9], the CMS does not diminish the effect of additive noise, and in some cases a combination of the CMS operation and RASTA processing can achieve better results than each of the methods separately.

First set of experiments presented in [9] focus on convolutive distortions such as those caused by the use of different microphones. In the logarithmic power spectrum these distortions have an additive constant effect, so the compressive transformation used was logarithmic. The band-pass filter was an IIR filter with the following transfer function:

$$H(z) = 0.1z^4 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}}. \tag{6}$$

This setup was used in an isolated digit recognition task, as well as in large vocabulary continuous speech experiments. In the continuous speech experiments, recognition performance of the PLP and RASTA-PLP methods were compared for clean speech as well as speech artificially filtered with a low-pass filter to simulate muffled speech. While the RASTA processing caused a slightly higher word error rate of 18.6 % when compared to the original 17.9 % in the clean speech case, for muffled speech the method achieved a word error rate of 33.0 %, less than half of the 67.5 % error rate of the unprocessed PLP features.

The second set of experiments described in [9] evaluated the speech recognition performance of the RASTA feature extraction method in an environment which had additive noise components. As additive noise manifests as a signal-dependent component in the logarithmic power spectrum, these experiments use an alternative compression function that has nearly linear behaviour for small spectral values, while still being logarithmic for large values. The compression and expansion functions used were:

$$y = \ln(1 + Jx) \qquad x = \frac{e^y}{J} \tag{7}$$

Here J is a signal-specific constant, and the expanding transformation is not the exact inverse, $x = \frac{e^y - 1}{J}$, of the compressive transformation, as it would not be guaranteed to be always positive, and would therefore necessitate some additional processing.

In these experiments, the optimal J value was found to be rather clearly dependant on the SNR value of the test data. Analysis of histograms of spectral energies suggest that the optimal choice for the J value is such that most of the noise falls into the linear part of the nonlinear compression function used, while most of the signal stays in the logarithmic part. With a suitable J value, RASTA processing was found to improve recognition rates at all noise levels. In isolated digit experiments, an automatically adaptive method of selecting J values achieved significantly improved results over the plain PLP method in environments with additive noise, convolutional distortions, or both, and also improvements over the purely logarithmic RASTA processing in the case of speech corrupted with additive noise.

4 CONCLUSIONS

The common mel-cepstral speech feature representation can be improved by using any of the more robust techniques for spectrum estimation. In the simplest case, the FFT based power spectrum estimate can be replaced with another estimate, as was done with a weighted linear prediction based spectrum in the SWLP-MFCC [7] method in section 3.1, and with a MVDR modeling based spectrum in the MVDR-MFCC [8] method in section 3.2.

A closely related alternative to MFCC features is the perceptual linear prediction (PLP) feature extraction process [6] outlined in section 3.1. PLP itself achieves some improvements on noise robustness over conventional MFCC features, but is not designed especially for noisy conditions. However, the RASTA processing introduced in section 3.3 is used in the RASTA-PLP feature extraction method [9] for the explicit purpose of reducing the effect of convolutional and additive noise. In addition, the MVDR-based PMCC [8] method introduced in section 3.2 closely resembles PLP processing in the way the parametric spectral modeling step is done for a perceptually filtered spectrum.

Both MFCC and PLP features incorporate perceptual considerations by filtering the raw spectrum estimate with a suitable filterbank, which additionally partially removes the undesired harmonic structure of speech. Because the MVDR estimation method is capable of tracking the upper spectral envelope and therefore also discards the harmonic structure, the PMVDR method [2] shown in section 3.2 avoids the filtering and applies the MVDR analysis on a directly warped spectrum.

Feature extraction methods that provide improved noise robustness can also be used together with other techniques with the same aim. [9] mentions the use of RASTA processing and CMS together, but in general changing the feature representation is a reasonably noninvasive process, so the efficiency of the different representations can in many cases easily be evaluated together with other tools.

5 REFERENCES

[1] S. B. David and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Trans. ASSP **28**, 357–366 (1980).

[2] U. H. Yapanel and J. H. L. Hansen, “A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition”, Speech Communication **50**, 142–152 (2008).

[3] B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music* (John Wiley & Sons, Inc., New York) (1999).

[4] J. Makhoul, “Linear prediction: A tutorial review”, Proc. IEEE **63**, 561–580 (1975).

[5] M. N. Murthi and B. D. Rao, “All-pole modeling of speech based on the minimum variance distortionless response spectrum”, IEEE Trans. Speech, Audio Process. **8**, 221–239 (2000).

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am. **87**, 1738–1752 (1990).

[7] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction", Speech Communication (2009), in press.

[8] S. Dharanipragada, U. H. Yapanel, and B. D. Rao, "Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method", IEEE Trans. Audio, Speech, Language Process. **15**, 224–234 (2007).

[9] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. Speech, Audio Process. **2**, 578–589 (1994).

Feature Enhancement in Automatic Speech Recognition

Jukka Pätynen

Helsinki University of Technology
Department of Media Technology
email: jukka.patynen@tml.hut.fi

Abstract

The purpose of feature enhancement is to remove degrading components from the speech signal and thus improve the recognition of speech in adverse conditions. In most approaches the nature of the speech signal, transmission path, or noise is simplified. The simplest methods are based on removing the estimated noise spectra from the signal. More complex techniques take advantage of the statistical estimates of the noise. A single microphone is applied in such techniques. Multiple microphones are used in advanced techniques based on more accurate noise estimation or beamforming. All methods apply statistical analysis on the signal at least to some extent, and they are differentiated mainly by their ability to handle non-stationary interfering components. This paper reviews several approaches on feature enhancement.

1 INTRODUCTION

Automatic speech recognition is principally based on identifying speech components from a sampled signal divided in short time frames. Sounds in the frames are classified with signal processing and statistical methods as phonemes that together in certain order form words, again constructing sentences. The analysis is performed by converting extracted acoustic features from the speech signal to phonemes and estimating the correct words based on the recognition history.

However, the recognition becomes unreliable when components other than the speech under recognition are added to the signal. This background noise can be almost anything, such as hum from computer or air conditioning nearby, traffic noise, a compressed-air drill, or another speaker.

The purpose of feature enhancement is to increase speech intelligibility in human-to-human communication or to improve the accuracy of automatic speech recognition by estimating the clean speech from the degraded signal. This makes feature enhancement an essential part of robust speech recognition techniques. Most enhancement methods rely heavily on statistical analysis, while more straightforward approaches exist as well.

Feature enhancement most often refers to reducing background noise from the signal. In the simplest cases the interfering components are assumed fairly stationary, since the recovering the original signal from the degraded recording without the knowledge of transmission path or interference is not possible.

This literature review is organized as follows. A basic model for speech recording is explained in the following section. After that we discuss various feature enhancement methods in two groups. First, the techniques utilizing a single microphone are presented.

Subsequently some multi-microphone applications are introduced. Finally, conclusions on the feature enhancement techniques are drawn. The structure follows loosely the overview paper by Compernolle [1].

2 BASIC SPEECH RECORDING MODEL

A straightforward model for recording speech is written as following equation

$$y[n] = h[n] * x[n] + d[n], \tag{1}$$

where h is the impulse response of the transmission path, x is the signal, and d is the additive noise at sampled moment n. Transmission path effect or channel filter $h[n]$ can be real-world phenomena such as room reverberation or low-quality transducers in communication devices. The equation can be applied in multitude of adverse conditions. A limitation of this model is that the transmission path is modeled with a linear filter. Therefore nonlinear phenomena, such as distortions, are not considered correctly. In real-world cases any of the variables is not known accurately.

In reality, the effect of the transmission path is also convolved with the additive noise, resulting in expression

$$y[n] = h[n] * (x[n] + d[n]). \tag{2}$$

However, we are not interested in the $d[n]$ in single microphone techniques, as the background noise is not available as *a priori* information. Thus, we can think the background noise only as $d[n]$. On the other hand, the human hearing system is not particularly sensitive to spectral changes or other effects resulting from the transmission path $h[n]$. This reduces the importance of compensating the channel filtering [1].

The generic model in Eq. (1) can be transformed to the frequency domain for methods utilizing the power spectrum:

$$Y(\omega) = X(\omega)\,|H(\omega)|^2 + D(\omega), \tag{3}$$

where $D(\omega)$ denotes the noise power spectrum. The power spectrum equation can be further transformed into another expression in the log-spectral domain [2].

$$\begin{aligned} y &= x + q + \log(1 + e^{d-x-q}) \\ &\approx x + f(x, d, q), \end{aligned} \tag{4}$$

where q represents an unknown linear channel filtering in log-spectral domain.

3 SINGLE MICROPHONE METHODS

3.1 Overlap add signal reconstruction

The overlap-add technique is not necessarily an exact feature enhancement technique, but it is rather a method for applying manipulation to the time domain signal. Processing is based on the assumption that the speech signal is nearly stationary in short time frames.

The degraded time-domain signal is windowed with window function, such as Hamming, having appropriate length. Each frame is transformed to frequency domain. The resulting magnitude response is modified with an arbitrary function and inverse-transformed back to time-domain. Modified window is then added to the result signal with pre-defined window overlap. Typical values for overlap are 2:1 or even 4:1 for decreased aliasing effects. The argument in the frequency domain is usually kept intact due to the relative hearing insensitivity to phase [1].

3.2 Spectral subtraction

Spectral subtraction is a straightforward method for suppressing stationary background noise. This technique is based on the overlap-add algorithm described in the previous section. The most obvious challenge in this approach is to get a good estimate of the noise power spectrum.

This method is a generic representation for noise suppression with spectral subtraction, which assumes an additional noise in the speech signal. However, the transmission path is not considered. Both speech and noise are regarded as uncorrelated random processes. Such processes imply that the power spectral densities are additive, when speech and noise are quasi-stationary in short time frame [1].

Each frame is transformed with FFT, where the total power spectrum is seen as a sum of clean signal and noise spectra

$$P_y(\omega) = P_x(\omega) + \hat{P}_d(\omega) \tag{5}$$

and by estimating $\hat{P}_d(\omega)$, a clean spectrum estimate is obtained:

$$\hat{P}_x(\omega) \approx P_y(\omega) - \hat{P}_d(\omega). \tag{6}$$

In essence, this assumption is valid only when spectral coefficients are independent and the signal window is infinitely long. With an additional assumption of signals consisting of a mixture of Gaussian distributions, the power spectral subtraction satisfies both maximum likelihood as well as least squares estimate $E[|\hat{P}_x(\omega) - P_x(\omega)|^2]$.

Two simple approaches exist to obtaining an estimate of the noise power spectra. They both rely on averaging short-time spectra in windowed frames that are assumed to contain only noise. Such frames can be taken either from the beginning of the degraded signal or identified with a speech detector.

Using the beginning of the transmission as noise reference is a trivial solution and robust for low SNR, but it does not provide any adaptive features i.e. in situation where the noise changes along the communication.

Applying the detection of the speech activity is a more complex alternative, but has the advantage of adaptivity. Compernolle [1] has described that for such implementation SNR should be at least 6 dB.

3.3 Wiener filtering

Wiener filtering is a method that in essence tries to create a filter that suppresses the additive noise from degraded signal minimizing the mean square error between clean signal estimate and original speech:

$$\hat{X}(\omega) = W(\omega)Y(\omega) \tag{7}$$

$$W(\omega) = \frac{\Phi_x(\omega)}{\Phi_y(\omega)}, \tag{8}$$

where the optimal Wiener filter would be $W(\omega)$.

In practice the optimal filter cannot be realized due to two problems. First, the optimal filter would usually be non-causal. Second, speech signal is not stationary in reality. Moreover, both speech and noise spectra should be known for optimal filter. To overcome these deficiencies the long-time noise spectra $P_d(\omega)$ is used with short-time speech spectra, which is approximated by $P_y(\omega) - P_d(\omega)$. This makes the Wiener filter somewhat comparable to the spectral subtraction approach [1].

3.4 Approach with vector Taylor series

This method is based on estimating the probability distribution function (PDF) of noisy speech given a clean speech PDF, a segment of noisy speech and a Taylor series expansion relating noisy speech to clean speech. Signal analysis is performed in log-spectral domain (see Eq. (3)). The additive noise in the observed signal is assumed to be near-Gaussian by its amplitude. Moreover, the log-spectra of the speech signal x is assumed to follow a sum of multivariate Gaussian distributions. The main idea behind the method is to approximate the function $f(x, d, q)$ in Eq. (4) with a vector Taylor series, where q represents the effect of unknown linear filtering in log-spectral domain.

The logarithmic spectrum of the speech signal is assumed to have a probability distribution from a summation of multivariate Gaussian distributions.

$$p(x) = \sum_{k=0}^{M-1} P[k] N_x(\mu_{x,k}, \Sigma_{x,k}) \tag{9}$$

In addition, noise is assumed to follow the statistic of a single Gaussian $N_d(\mu_d, \Sigma_d)$. In order to have a suitable compensation, variables for unknown channel filtering as well as noise mean and variance have to be determined. Moreover, probability distribution for y has to be computed given $p(x)$, μ_d and Σ_d. Although this is not a trivial problem, analytical solution is possible with the vector Taylor series approach[2]. The Taylor series expansion gives an accurate representation around the mean of a Gaussian even with few terms from the Taylor series. The vector Taylor series approximation for $f(x, d, q)$ is

$$\begin{aligned} f(x, d, q) \approx & f(x_0, d_0, q_0) + \frac{d}{dx} f(x_0, d_0, q_0)(x - x_0) + \\ & \frac{d}{dd} f(x_0, d_0, q_0)(d - d_0) + \frac{d}{dq} f(x_0, d_0, q_0)(q - q_0) + \dots . \end{aligned} \tag{10}$$

Moreno *et al.* [2] have used two variants of the Taylor series approximation where either terms up to zeroth of first order are utilized. In the latter version the terms shown in Eq. (10) are applied. In the zero-order vector Taylor series expansion for $f(x, d, q)$ gives a Gaussian distribution $p(y) = N_y(\mu_y, \Sigma_y)$ for noisy speech y with following mean vector and covariance matrix:

$$\mu_y = E(y) = E(x + f(x_0, d_0, q_0)) = \mu_x + f(x_0, d_0, q_0) \quad \text{, and} \tag{11}$$

$$\Sigma_y = \Sigma_x. \tag{12}$$

For the first-order Taylor series these parameters are computed analogously to Eq. (10). With both zero- and first-order expansions, q and μ_d are solved iteratively, and it is described by Moreno [2] and Dempster [3].

The actual compensation for noisy speech is estimated by minimum mean square error (MMSE) method for obtaining clean speech given the noisy speech

$$\hat{x} = E_x(x|y) = \int x p(x|y) dx \tag{13}$$

$$\hat{x} = \int (y - f(x, d, q)) p(x|y) dx \tag{14}$$

$$= y - \sum_k p(k|y) f(\mu_{x,k}, \mu_d, q) \quad \text{(zeroth-order)}. \tag{15}$$

3.4.1 Particle filter approach to vector Taylor series

The vector Taylor series approximation of the log-spectral features of the clean speech has been improved by Faubel and Wölfel [4]. In this method particle filters (PF) are applied in order to remove the need for Taylor series approximation. Here, this technique is discussed only very briefly.

Previously, the effect of the environment to the speech spectrum was modeled with a low-order Taylor series. In particle filter method, inferring the clean speech from noisy signal is expressed as tracking the clean speech spectrum x_t from a history of noisy spectra $y_{1:t}$ by using the conditional probability $p(x_t|y_{1:t})$. A solution for such problem is to find the conditional mean $E[x_{1:t}|y_{1:t}]$. The evolving model of the noise spectrum is sampled with particles $j = 1 \ldots N$ having individual importance weights. In the probability distribution the modeled noise is introduced as hidden variable in the marginal distributions $p(x_t, d_t|y_{1:t})$. Normally the noise would not be present in the equations due to integration over the variable [5].

$$E[x_t|y_{1:t}] = \int \underbrace{\int x_t \cdot p(x_t|y_{1:t}, d_t) dx_t}_{=:h_t(d_t)} p(d_t|y_{1:t}) dd_t \tag{16}$$

$$\approx \sum_{j=1}^{N} h_t(d_t^{(j)}) \widetilde{w}_t^{(j)} \quad , \tag{17}$$

where $\widetilde{w}_t^{(j)}$ is a normalized importance weight vector for noise hypothesis $h_t(d_t^{(j)})$. Conse-

quently, $h_t(d_t)$ is expressed as sum of K separate Gaussians in the clean speech spectrum:

$$h_t(d_t) = \sum_{k=1}^{K} p(k|y_{1:t}, d_t) \int p(x_t|k, y_{1:t}, d_t) dx_t. \tag{18}$$

In the spectral domain the noise is considered to shift the means of the clean speech distribution $p(x_t)$ for kth Gaussian:

$$e^{\mu'_k} = e^{\mu_k} + e^{\mu_t} \Leftrightarrow \mu'_k = \mu_k + \underbrace{\log(1 + e^{d_t - \mu_k})}_{=:\Delta_{\mu_k, d_t}}. \tag{19}$$

Clean speech spectrum can be obtained from spectrum y_t and the estimated mean shift:

$$x_t = y_t - \Delta_{\mu_k, d_t}. \tag{20}$$

This technique was been noticed to slightly improve word error rates compared to vector Taylor series method. However, the computational load is claimed to be lower with particle filter approach [4].

3.5 Stereo-based piecewise linear compensation for environments (SPLICE)

Basic methods for noise suppression usually assume a model for the source of distortion or degradation to the signal. The approach used in SPLICE (stereo-based piecewise linear compensation for environments) manages to overcome this deficiency by not making any assumptions on how the noisy signal is produced from the clean signal. The advantage with this technique is that noise removal is not restricted to any separate distortion. While other methods form functions based on modeled interferences, SPLICE can model combinations of several simultaneous distortions. In time domain, SPLICE functions on short signal frames, and statistical analysis is performed in cepstral domain, (e.g. FFT of FFT of signal frame).

In SPLICE, each distortion type is trained from noisy data to establish a probability mixture model $p(\mathbf{y})$ for noisy speech $\mathbf{y}$ against clear speech $\mathbf{x}$ (hence, stereo-based). Typically the mixture consists of 256 Gaussian components. A joint probability distribution is expressed as follows:

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) \tag{21}$$

However, $p(\mathbf{x}|\mathbf{y})$ usually includes some parameters that are nonlinear functions of $\mathbf{y}$. Therefore an additional variable s is used to divide the probability space into regions, where $\mathbf{x}$ and $\mathbf{y}$ are in approximately linear relation. Now, Eq. (21) has the form

$$p(\mathbf{x}, \mathbf{y}) = \sum_{s} p(\mathbf{x}|\mathbf{y}, s)p(\mathbf{y}|s)p(s), \tag{22}$$

where $\mathbf{x}$ and $\mathbf{y}$ are approximately linear. Additionally,

$$p(\mathbf{x}|\mathbf{y}, s) = N(\mathbf{x}; \mathbf{y} + \mathbf{r}_s, \Gamma_s) \quad \text{and} \tag{23}$$

$$p(\mathbf{y}, s) = N(\mathbf{y}; \mu_s, \Sigma_s). \tag{24}$$

A vector $\mathbf{r}_s$ is generated for estimating the corrected speech on the basis of the structure of the noisy cepstrum in the training phase with maximum likelihood criterion. Eq. (23) indicates the probability with variance Γ_s for clean speech estimate on condition of the noisy speech with parameter s and $\mathbf{r}_s$.

The estimated cepstrum for clean speech is conditionally obtained from basis of the degraded speech:

$$\begin{aligned}\hat{\mathbf{x}} &= E_x[\mathbf{x}|\mathbf{y}] \\ &= \sum_s p(s|\mathbf{y}) E_x[\mathbf{x}|\mathbf{y}, s] \\ &= \mathbf{y} + \sum_s p(s|\mathbf{y})\mathbf{r}_s.\end{aligned} \tag{25}$$

This gives the minimum mean square error estimate cepstrum vector of the clean speech as a linearly weighted sum of the separate correction vectors and the degraded signal.

4 MULTIPLE MICROPHONE METHODS

4.1 Adaptive noise canceling

Least mean squares (LMS) algorithm is based on an adaptive filter which is automatically adjusted to suppress interference contained in a recorded signal. Unknown background noise is transmitted into the recorded speech signal with unknown path. The filter tries to match itself to the impulse response of the background noise transmission path. [1]

The algorithm relies on having a noise reference without speech signal. Therefore an additional microphone is required for capturing only the noise in the speech environment. For this reason LMS algorithm is not directly comparable to techniques not needing any quasi- *a priori* information about the recorded signal. The advantage here is that stationary as well as non-stationary interference can be suppressed with the continuously monitored reference noise.

In theory, two signals are captured, where the first one contains speech with additive noise convolved with unknown impulse response and the second one only noise.

$$y_1[n] = x[n] + h_1[n] * d[n] \tag{26}$$

$$y_2[n] = h_2[n] * d[n] \simeq d[n] \tag{27}$$

One should note that the notation resembles to the the recording model in Eq. (2). The impulse response $h_2[n]$ in the second equation is omitted, as it cannot be resolved.

The adaptation in the method is performed by having a filter $w^n[n]$ based on the knowledge on the noise $y_2[n]$ and the degraded signal $y_1[n]$. With appropriate update parameters the impulse response of the filter converges to the impulse response of the unknown noise transmission path. As the result, correlated additive noise in $y_1[n]$ is suppressed. It is noticeable that the estimate the transmission path for speech is not

considered in these equations.

$$\hat{x}[n] = y_1[n] - w^n[n] * y_2[n] \tag{28}$$

$$w^{n+1}[l] = w^n[l] + 2\mu\hat{x}[n]y_2[n-l] \tag{29}$$

Higher values of constant μ provide faster rate of adaptation. However, there is a certain value that should not be exceeded with regard to the filter length L and power spectra of recorded signals in order to ensure the stability of the filter.

$$\mu < \frac{1}{10L} P_{y_2} \tag{30}$$

With a poor signal-to-noise ratio the adaptation constant should be evaluated against the recorded signal having higher power. Besides the sufficiently low μ, it is vital to have $y_2[n]$ completely separated from the speech. Otherwise the algorithm tries to suppress also the speech, leading to distortion in the output signal. Two concepts concerning the signal levels should be considered [1]:

1. Signal should leak less to the noise measurement than the noise to signal measurement, or
2. the noise source power should be much larger than the signal source power.

These limitations makes the adaptive noise control technique applicable for conditions where the noise level is particularly high, such as loud factories or airplanes.

4.2 Beamforming

Beamforming methods differ from the adaptive noise canceling by utilizing multiple degraded signal recordings instead of separate noise reference. In addition, many algorithms rely on the known geometry of the microphone. Beamforming is particularly applicable in situations where the signal has to be precisely captured in a specific position from a distance in noisy conditions. Beamforming can be implemented in various ways, of which two of the most common are briefly presented here.

4.2.1 Delay and sum beamforming

The main principle in basic beamforming is to use multiple recorded signals and, apply delays to them, and finally sum the delayed signals. In the simplest case of two microphones in plane with the speech source, the situation is backwards analogous to constructive and destructive interference from two sound sources (e.g., microphones). By appropriately delaying another of the source signals the receiver detects amplified or attenuated signal. In case of beamforming the delays are applied to microphones. The principle can be

formulated as follows.

$$y_m[n] = x[n - \tau_m] + d_m[n] \quad \text{(delay)} \tag{31}$$

$$\hat{x}[n] = \frac{1}{M} \sum_{m=1}^{M} y_m[n + \hat{\tau}_m] \quad \text{(sum)} \tag{32}$$

The delays $\hat{\tau}_m$ for each microphone signals can be estimated with cross-correlation combined to the microphone geometry [6]. Advantages with delay and sum beamforming is the robustness, while errors in the estimated delays cause only little distortion. However, the gained SNR increases slowly with higher number of microphones.

4.2.2 Constrained adaptive beamforming

The constrained adaptive beamforming is an improved method based on the basic technique. An adaptive noise canceling is applied to the beamformer structure by utilizing difference of the delayed signals as noise reference. The noise reference is then filtered with adaptive filter and finally subtracted from the basic beamformed speech signal. The resulting signal adjusts the adaptive noise reference filter, such as a Wiener filter. [7]

4.3 Blind signal separation

Blind signal separation is based on assumed crosstalk mixtures between several recorded signals. The signals are also assumed as statistically independent, e.g., separate speakers. The recorded signals are closely related to adaptive noise canceling in Eqs. (27) and (29):

$$y_1[n] = x_1[n] + h_1[h] * x_2[n] \tag{33}$$

$$y_2[n] = x_2[n] + h_2[h] * x_1[n], \tag{34}$$

where $\hat{s}_1$ and $\hat{s}_2$ are obtained by estimating transfer functions $h_m[h]$ as presented in Eq. (29).

Statistics of the mixed signals can be used to improve the source separation with blind techniques. Such aspects are discussed more thoroughly by Cardoso [8] or Der [9].

5 CONCLUSIONS

In this paper some common approaches on speech feature enhancement have been discussed. The simplest methods apply windowing time domain signal to short time frames and traditional filtering. More advanced techniques such as vector Taylor series are based on signal statistical analysis.

6 REFERENCES

[1] D. V. Compernolle, "DSP techniques for speech enhancement", in *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, 21–30 (1992).

[2] B. J. Moreno, B. Raj., and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 733–736 (1996).

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm", J. of the Royal Stat. Soc. **39**, 1–38 (1977).

[4] F. Faubel and M. Wölfel, "Overcoming the vector taylor series approximation in speech feature enhancement – a particle filter approach", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 557–560 (2007).

[5] F. Faubel and M. Wölfel, *Speech feature enhancement for speech recognition by sequential Monte Carlo methods, Diploma thesis* (Universität Karlsruhe, Germany) (2006).

[6] H. F. Silverman, "Some analysis of microphone arrays for speech data acquisition", IEEE Transactions on Acoustics, Speech and Signal Processing **35**, 1699–1712 (1987).

[7] S. Nordholm, I. Claesson, and P. Eriksson, "The broad-band wiener solution for griffiths-jim beamformers", IEEE Transactions on Signal Processing **40**, 474–478 (1992).

[8] J. F. Cardoso, "Blind signal separation: statistical principles", Proc. IEEE **9**, 2009–2025 (1998).

[9] R. Der, *Blind signal separation* (Report of Dept. of Electrical and Computer Engineering, McGill University) (2001).

Model Compensation for Noise Robust Speech Recognition

Peter Smit

Helsinki University of Technology
Department of Information and Computer Science
email: peter@smitmail.eu

Abstract

Ideally we would like to have our speech recognition models trained for the environment where they are going to be used in. However, different models are most of the time not feasible, because recording and noise conditions are changing. Retraining costs a lot of time and resources. Therefore, methods are developed to adjust clean speech models to fit the current environment. One of the methods is called model compensation.

The basic idea is to create a separate hidden Markov model (HMM) of the noise which compensates the trained HMM of clean speech. The separate HMM of the noise is a small model (as it has a few states) and can be estimated at a speech pause. After the estimation, the both models are converted into the domain linear in power where they can be easily combined together. This is performed under the assumption that the noise and speech are additive and independent. The combined model in the linear domain will be converted back to the cepstral domain and used instead of the pure clean speech model in speech recognition.

Most of the composition techniques are based on Parallel Model Combination. This technique give estimations for static parameters of models. For the estimation of dynamic parameters Data driven PMC is discussed.

Also for convolutional noise models can be compensated. HMM decomposition is providing an approximation of the acoustic transfer model.

1 INTRODUCTION

There has been research in speech recognition for multiple decades. Still, it is a hard task to recognize continuous speech, especially when the recording is subject to a noisy environment or poor recording equipment like phones.

The base of the most speech recognition methods are meant for clean speech. To make these methods more noise robust multiple methods are developed.

In this article mainly Parallel Model Combination (PMC) [1] will be discussed. The basic variant of this method is limited in the variation of noise, the type of noise, and in the parameters used (only static parameters). Other methods that will be discussed are mainly extensions to PMC to overcome some of these limitations.

2 MODEL COMPENSATION

The purpose of model compensation is to replace the HMM used in a speech recognition with a compensated one. This compensated model is a combined model of the trained cleaned speech HMM and a noise model.

The basic assumption of Model Compensation is that noise is additive in the linear power domain. However, extensions are made to also support convolutional noise. Additive noise means that noise will not modify the signal of the speech itself, except for adding something to it. Theoretically it would be not an ideal method to recognize degraded speech (for example in phone lines), but even there improvements can be seen in comparison to the standard model.

To be able to combine a noise HMM and the clean speech HMM, the models must be converted back to the linear domain. A HMM normally operates on the cepstral domain. The parameters that have to be transformed back are the means and variances of the mixture models of the HMM. The transition weights between states will stay the same.

In the linear domain the values can be added together (with a weight for different speech/noise ratios). After that the parameters can be converted back to the cepstral domain.

If the noise model has multiple states, the resulting HMM will be having combined states of noise and phonemes and the transitions between them can be estimated by the product of the transition weight between noise states and the transition weights between phoneme states.

3 PARALLEL MODEL COMBINATION

3.1 Basic Parallel Model Combination

Parallel Model Combination (PMC) is a method developed at the university of Cambridge by Gales and Young [1].

The base of PMC is the mismatch function describing the relation between noise and speech. The standard mismatch function is

$$O(t) = gS(t) + N(t) \tag{1}$$

where O is the corrupted speech signal, S is the clear speech and N is the noise. This variables are expressed in the linear domain and are all varying over time. g is a term to correct for different ratios between speech and noise. The mismatch function is based on the assumptions that the noise and speech are independent, additive and that a Gaussian mixture model can represent the vectors in the cepstral domain.

In the cepstral domain there is a noise HMM with one state and a clean speech HMM. The clean speech HMM is already trained and the noise HMM can be trained in a speech pause. The models can be both described by their mean and variances ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ for each state). The target is to calculate the means and variances of a new noisy speech model $\{\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i\}$ from the mean and variance of the noise model $\{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\}$ and the means and variances of the clean speech model $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$. The name Parallel Model Combination comes from the fact that the Noisy Speech Model is based on two parallel working models; the noise and the clean speech one.

The combination of the models is not trivial. The models are trained in the cepstral domain, while the mismatch function is defined in the linear domain. Therefore the

parameters of the models have to be mapped back to the linear domain first where they can be combined and after that be converted to the cepstral domain to be used in the noisy speech model.

The means and variances are mapped back to the linear domain in two steps. First a transformation is done to the log domain and secondly the parameters are estimated in the linear domain.

To map the parameters to the log-spectral domain an inverse of the applied cosine function is done. The parameters are multiplied by the inverse of matrix C.

$$\boldsymbol{\mu}^l = \boldsymbol{C}^{-1}\boldsymbol{\mu}^c \tag{2}$$

$$\boldsymbol{\Sigma}^l = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}^c \left(\boldsymbol{C}^{-1}\right)^T \tag{3}$$

In these equations the superscript c indicates the cepstral domain and the superscript l the log spectrum domain. For the domain linear in power no superscript will be used.

For the mapping from the log domain to the linear domain the parameters can be estimated by taking the expectation. This because the parameters represent (Gaussian) probability models. When all the distributions are Gaussian and only static parameters are used (no delta and delta-delta parameters), the formulas for the means and variances can be derived to [1, 2]

$$\mu_i = e^{\mu_i^l + \Sigma_{ii}^l/2} \tag{4}$$

$$\Sigma_{ij} = \mu_i \mu_j \left(e^{\Sigma_{ij}^l} - 1\right) \tag{5}$$

The means and variances of the noise and clean speech model can now be added together, with the clean speech multiplied with a gain matching term g to take in account that the levels of speech and noise can variate. The resulting equations are

$$\hat{\boldsymbol{\mu}} = g\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}} \tag{6}$$

$$\hat{\boldsymbol{\Sigma}} = g^2\boldsymbol{\Sigma} + \tilde{\boldsymbol{\Sigma}} \tag{7}$$

The gain function can vary. In a environment with stationary noise the following function could be used

$$g = \frac{E_{ns} - E_n}{Es} \tag{8}$$

In this function E_{ns} is the energy of the noisy speech, E_n the energy of the noise and E_s the energy of the clean speech.

Now the models are combined, the parameters can be mapped to the cepstral domain. From the linear to the log domain the formulas are

$$\hat{\mu}_i^l = \log\left(\hat{\mu}_i\right) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1\right) \tag{9}$$

$$\hat{\Sigma}_{ij}^l = \log\left(\frac{\hat{\Sigma}_{ij}}{\hat{\mu}_i\hat{\mu}_j} + 1\right) \tag{10}$$

And after that to the cepstral domain.

$$\hat{\boldsymbol{\mu}}^c = \boldsymbol{C}\hat{\boldsymbol{\mu}}^l \tag{11}$$

$$\hat{\boldsymbol{\Sigma}}^c = \boldsymbol{C}\hat{\boldsymbol{\Sigma}}^l\boldsymbol{C}^T \tag{12}$$

The means and variances are now used for the mixture models in the states of the new HMM.

The new noisy speech HMM does have the same amount of states as the clean speech HMM. Every state is a combination of a state from the clean speech HMM and the state of the noise HMM. The transition probabilities are not influenced by the noise and are therefore the same as in the clean speech HMM.

3.2 PCM with multiple noise states

With basic PCM the assumption is made that there is only one state in the noise HMM. However, for non stationary noise, often a multi state noise HMM will give a better performance [1].

If the multiple noise states would be combined with the HMM this would normally need recalculation of the model taking in consideration the joint probabilities between all states. However, when the noise model is fully connected, the calculation is easier. In that case the noisy speech HMM can be constructed by taking the full product of the both HMM's (combining each possible noise and clean speech state) and assign as transition probabilities the product of the transitions probabilities in the noise model and in the clean speech model.

The resulting model will contain M x N states, where M is the number of states in the clean speech model and N the number of states in the noise model. Normally N will be quite small (like 2 or 3) and will therefore not increase the computational load of the recognition system too much.

3.3 PCM with Delta and Delta-Delta parameters

In basic PCM only the basic static parameters are considered. In state-of-the-art systems however, also delta and delta-delta parameters are used. To combine a noise and clean speech model which have these parameters is more complicated that than the standard mapping in basic PCM. [3, 4]

The delta and delta-delta parameters are used for taking the correlation between successive frames into account. However to find the correlation between frames, full or at least partially full covariance matrices should be used. This is not desirable because of increase storage requirements.

These statistics can also be expressed in the statistics already known. However to take the expectation of the delta parameters is not possible in a closed form solution. These equations are too extensive for this review but can be found in [2].

Because no closed form solution is available, the parameters must be approximated,

Model	Test set	Word Err
Clean	Feb89	38.7
	Oct89	32.0
	Feb91	33.4
Noisy	Feb89	7.3
	Oct89	8.6
	Feb91	6.9
PMC	Feb89	8.3
	Oct89	8.1
	Feb91	7.3

Table 1: Comparison of different models [3]

for example by numerical integration. Numerical integration is costing a lot of resources and makes this method less used in practice.

3.4 Data driven PCM

As an alternative to approximating the delta and delta-delta parameters there are other methods which take higher order statistics into account.

Data driven PCM is a method where the noise model and the clean speech model are used to generate virtual samples of noisy speech [3]. This samples are then used to calculate a Maximum likelihood estimate for the noisy speech HMM. This is a computationally easier task because no integrals have to be computed have to be done. Still all features, so also the dynamic parameters, are taken into account. The effectiveness depends on the number of data samples generated.

3.5 Results

PMC is able to come close to the word error rate of a model trained on noisy speech. In comparison to a model with no noise robust properties an improvement of 75% can be made. In table 1 results for three different test sets and the three different models is shown [3]. The data has 18-20dB Lynx helicopter noise.

4 CONVOLUTIONAL NOISE

In PMC only additive noise is considered. There are extensions to take in account also convolutional noise. Convolutional noise are transformations of the clean speech like channel distortion, reverberation, or microphone characteristics. One of the methods to take convolutional noise in account is HMM Composition for reverberant speech [5]. This method works the same as PMC except that the instead of the standard PCM mismatch equation (1) a new mismatch equation is used.

$$O(t) = H(t) \cdot S(t) + N(t) \tag{13}$$

Where H is an acoustic transfer function. This function is can be depended on time. A person could walk away from the microphone, for example, and therewith change the effect of the acoustic transfer function. In practice however, often a static transformation matrix is used.

Because the combination of the clean speech and the acoustic transfer model are a product in the linear domain, it is a simple addition in the cepstral domain [6]. The problem is the estimation of this acoustic transfer model.

4.1 HMM Decomposition for reverberant speech

HMM Decomposition gives a method for the estimation of the acoustic transfer model for e.g. reverberant rooms [6]. The principle is that from an estimated complete model, which is assumed to be estimated for speech in noisy reverberant conditions, the noise element and the clean speech element are removed by decomposition thus leaving the transfer model. The speech element is removed in the cepstral domain and the noise element is removed in the linear domain.

This method needs to estimate a model for the noisy transformed speech, by adapting the current noisy speech model. The resulting model will contain all the effects of the 3 separate models. The results of this method are depended on the adaption method and the amount of adaption data used.

5 CONCLUSIONS

Model compensation and especially basic PMC is able to make the speech recognition system more robust effective and resource cheap way.

To obtain also good results when dynamic parameters are used there are different options. Numerical integration and the Monte-Carlo sampling technique DPMC. Of these two the latter is computationally less heavy.

For convolutional noise HMM decomposition can provide in a good method to approximate the acoustic transfer function model.

6 REFERENCES

[1] M. J. F. Gales and S. J. Young, “Cepstral parameter compensation for HMM recognition in noise”, Speech Communication **12**, 231–239 (1993).

[2] M. J. F. Gales, “Model-based techniques for noise robust speech recognition”, Ph.D. thesis, University of Cambridge (1995).

[3] M. J. F. Gales and S. J. Young, “A fast and flexible implementation of parallel model combination”, in *1995 International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95.*, volume 1 (1995).

[4] M. J. F. Gales and S. J. Young, “Robust continuous speech recognition using parallel

modelcombination", IEEE Transactions on Speech and Audio Processing **4**, 352–359 (1996).

[5] S. Nakamura, T. Takiguchi, and K. Shikano, "Noise and room acoustics distorted speech recognition by HMMcomposition", in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings.*, volume 1 (1996).

[6] T. Takiguchi, S. Nakamura, Q. Hou, and K. Shikano, "Model adaptation based on HMM decomposition for reverberant speechrecognition", in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97.*, volume 2 (1997).

Voice Activity Detection with Speaker Localization Application

Teemu Korhonen

Tampere University of Technology
Department of Signal Processing
email: teemu.korhonen@tut.fi

Abstract

The presence of a human speaker can be detected using methods often classified under the term of *Voice Activity Detection* (VAD). Ideally the detection scheme should separate the time instances with active voice of a speaker from the ones without voice. The activity information can be used for preprocessing of the data, separating frames of interest from others, but more advanced algorithms integrate the activity information to the estimation process itself. In this literature review different methods for voice activity detection are presented. Also, an acoustic source localization application using VAD is presented as an example of an interesting way of taking advantage of the extra information.

1 INTRODUCTION

Human voice is a transient phenomenon in nature. Transience becomes obvious even with a single sentence which comprises of multiple time separate parts. Spoken word is riddled with moments of silence (or background noise) caused by syllable and word separations, pronunciation and pauses of deliberation. The discontinuity is especially harmful when an algorithm assumes the speech data to be continuous: for the acoustic speaker localization this would mean association of a speaker position even to a silence. When the data does not match the assumptions, any following steps of the algorithm are effectively processed using noise as input and the results become erroneous.

Following sections deal with different methods of voice activity detection (VAD), shortly presenting the relevant theory and operational issues. Section 2 presents basics behind VAD and couple of algorithms chosen for their robustness against adverse noise conditions. In section 3 the VAD is presented as audio segmentation (sub)problem. Section 4 deals with an application of the VAD within an acoustic source localization framework, where voice activity information has been used to enhance the performance of a speaker tracking algorithm.

2 VOICE ACTIVITY DETECTION

In this section a group VAD algorithms is presented with high separation rates between speech/non-speech [1]. The VAD algorithms are based on a feature vector $\mathbf{x}$ extracted from the signal $x(n)$ and assume that the speech signal is corrupted by additive noise. The basic problem of the VAD can be formulated as hypothesis test between scenarios

where there is only noise and the one with both noise and speech - more formally:

$$H_0 : \mathbf{x} = \mathbf{n}$$
$$H_1 : \mathbf{x} = \mathbf{n} + \mathbf{s},$$

where H_1 corresponds to the scenario with speech and H_0 is the null-hypothesis "only noise". The posterior probability of the hypotheses given observation of feature $\mathbf{x}$ should follow

$$P(H_1|\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} P(H_0|\mathbf{x}). \tag{1}$$

That is, the hypothesis with larger posterior probability is selected as the winner. The equation can be modified using the Bayes rule ($P(y|x)P(x) = P(x|y)P(y)$) to yield a likelihood ratio test:

$$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P(H_0)}{P(H_1)}, \tag{2}$$

which now decides between the two possible hypotheses defined before. Division of the signal $x(n)$ in discrete Fourier transform coefficients $j = 0, \cdots, J-1$ of the speech (S_j) and noise (N_j) yields a multi-Gaussian parameterization of the signal and a solution:

$$p(x|H_0) = \prod_{j=0}^{J-1} \frac{1}{\pi\sigma_N^2(j)} \exp\{-\frac{|X_j|^2}{\sigma_N^2(j)}\} \tag{3}$$

$$p(x|H_1) = \prod_{j=0}^{J-1} \frac{1}{\pi\sigma_N^2(j) + \sigma_S^2(j)} \exp\{-\frac{|X_j|^2}{\sigma_N^2(j) + \sigma_S^2(j)}\} \tag{4}$$

where X_j is the j-th bin noisy speech DFT coefficient $\sigma_N^2(j)$ and $\sigma_S^2(j)$ stand for binwise variances for noise and signal, respectively. Applying the Bayesian hypothesis test given in (2) the equation is reduced to

$$\frac{1}{J}\sum_{j=0}^{J-1}\left[\frac{\gamma_j\xi_j}{1+\xi_j} - \log(1+\gamma_j)\right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \quad \xi_j = \frac{\sigma_S^2(j)}{\sigma_N^2(j)}, \gamma_j = \frac{|X_j|^2}{\sigma_N^2(j)}, \tag{5}$$

where the η defines the decision threshold and ξ_j and γ_j stand for *a priori* and *a posteriori* SNR of the signal. These SNR-coefficients can be estimated using the Ephraim and Malah minimum square-error (MMSE) estimator [2]. The threshold η must also be trained e.g. using noise sequences as training data [3].

In speech production process word beginnings and endings have often reduced signal energy. This motivates VAD decision postprocessing for removal of outliers using e.g. smoothing. These so called hang-over algorithms extend and smooth the VAD decision in order to better manage speech periods masked by the acoustic noise.

2.1 Long-term spectral divergence

Noisy speech signal $x(n)$ is processed in overlapping frames and $X(k, l)$ stands for the amplitude spectrum for k-th band of frame l. The N-order Long-Term Spectral Envelope

(LTSE) is defined as:

$$LTSE_N(k,l) = \max\{X(k,l+j)\}_{j=-N}^{j=+N}. \tag{6}$$

The VAD decision per each frame l is formulated as speech and noise ratio using average noise spectrum magnitude $N(k)$ for each band $k = 0, 1, \ldots, NFFT-1$ and the LTSE as

$$LTSD_N(l) = 10\log_{10}\left(\frac{1}{NFFT}\sum_{k=0}^{NFFT-1}\frac{LTSE^2(k,l)}{N^2(k)}\right)\begin{matrix}H_1\\>\\<\\H_0\end{matrix}\eta. \tag{7}$$

2.2 Multiple observation likelihood ratio test

Multiple observation likelihood ratio test (MO-LRT) uses larger range of observations to render the decision. The range of features vectors used is within m frames of the frame of interest l and can be presented as vector: $\{\mathbf{x}_{l-m}, \ldots, \mathbf{x}_l, \ldots, \mathbf{x}_{l+m}\}$. The MO-LRT is sum of the log likelihood ratios over the range:

$$l_{l,m} = \sum_{k=l-m}^{l+m}\ln\left(\frac{p(\mathbf{x}_k|H_1)}{p(\mathbf{x}_k|H_0)}\right)\begin{matrix}H_1\\>\\<\\H_0\end{matrix}\eta. \tag{8}$$

The use of frames around the current one postpones the decision for m frames, but gives a significant improvement in speech/non-speech discrimination accuracy without need of a separate decision smoothing.

2.3 Order statistics filter

The Order Statistics Filter (OSF) is defined for l-th frame as decision in terms of the average subband SNR using N frames around the frame of analysis. The decision for OSF is

$$SNR(l) = \frac{1}{K}\sum_{k=0}^{K-1}QSNR(k,l)\begin{matrix}H_1\\\geq\\<\\H_0\end{matrix}\eta, \tag{9}$$

where subband SNR is defined as difference

$$QSNR(k,l) = Q_p(k,l) - E_N(k), \tag{10}$$

and the components:

$$E_N(k) = \alpha E_N(k) + (1-\alpha)Q_{0.5}(k,l), \tag{11}$$

$$Q_p(k,L) = (1-f)E_{(s)}(k,l) + fE_{(s+1)}(k,l). \tag{12}$$

The $Q_p(k,l)$ is the sampling quantile using $s = \lfloor 2pN \rfloor$ and $f = 2pN - s$ and the quantile parameter $p \in [0,1]$ (setting $p = 0.5$ it equals a median filter). $E_N(k)$ is the noise level of k-th band, that is updated due to non-stationary noise assumptions using α as ratio between old value $E_N(k)$ and the sampling quantile. $E_{(r)}(k,l)$ stands for r-th largest number in algebraic order (value sorting) from a sequence of log-energy values around the

current frame l: $\{E(k, l-N), \ldots, E(k, l), \ldots, E(k, l+N)\}$.

Finally the log-energies $E(k, l)$ are computed for l-th frame and k-th subbands:

$$E(k,l) = \log\left(\frac{K}{NFFT}\sum_{m=m_k}^{m_{k-1}-1}|Y(m,l)|^2\right), m_k = \lfloor\frac{NFFT}{2K}k\rfloor, \tag{13}$$

with K subbands $k = 0, 1, \ldots, K-1$. The $Y(m, l)$ stands for Wiener de-noised input signal for m-th band at frame l. For an extensive description of the algorithm see the work Ramirez *et al* [4].

3 AUDIO SEGMENTATION

Voice activity detection can also been seen as a subproblem to audio segmentation, where each part of the audio data receives a label indicating nature of the source. When the set of possible segments becomes a binary decision between speech and non-speech the segmentation is effectively a VAD.

Audio segmentation has been used in automatic classification of historical speeches by Hansen *et al.* [5]. The audio data is processed in two parts: a coarse segmentation is generated using a distance measure in a 45-dimensional feature space and then compensated against false alarms using a distance measure in a feature space from Principal Component Analysis (PCA). Besides speech activity, the *CompSeg* method also does classification to music, commercials, background noise and other events. A list of features used by the algorithm:

- **PMVDR:** Perceptual Minimum Variance Distortionless Response (12 static + 11 delta-PMVDR + 1 static energy) [6, 7]
- **FBLC:** FilterBank Log energy Coefficients (20 Mel frequency coefficients)
- **SZCR:** Smoothed Zero Crossing Rate (1 coefficient from 5-point mean ZCR) [8]
- **PCA:** Principal Component Analysis of the features (22 first components) [9]

The segmentation is done using distance measure depending on the segment length. Shorter segments (below 5 seconds) are classified using a Hotelling T^2-statistics:

$$T^2 = \frac{ab}{a+b}(\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1}(\mu_1 - \mu_2), \tag{14}$$

while longer segments are processed with a faster Bayesian Information Criterion (BIC). The weighted mean distance of the block diagram is a T^2-statistics with identity matrix as covariance.

4 APPLICATION IN ACOUSTIC SOURCE LOCALIZATION

Acoustic source localization is a problem field concentrating on finding a speaker position under adverse conditions, such as environmental noise and reverberation. The speech

signal is seldom continuous and the gaps caused by silence misguide the tracking algorithms trying to pinpoint the speaker position. The Voice Activity Detection framework could be applied as postprocessing with removal of the frames without any speaker. More advanced method proposed by Lehmann and Johansson [10, 11, 12] integrates the VAD information to the tracking process.

The basis of the algorithm is Steered BeamForming (SBF) which gives spatial amplification to a point l as a sum over delayed versions of signals of M spatially separate microphones:

$$P(l) = \int_{\Omega} |\sum_{m=1}^{M} W_m(\omega) F_m \mathrm{e}^{j\omega \|l - l_m\|/c}|^2 \mathrm{d}\omega, \tag{15}$$

where l_m denotes a m-th microphone position, c the speed of sound, $W_m(\cdot)$ frequency weighting term (often phase transform) and $F_m(\cdot) = \mathcal{F}\{f_m(t)\}$ is the Fourier transform of the signal $f_m(t)$. The frequency range of interest Ω is typically between $300Hz$ and $3000Hz$ for speech processing.

A local maximum in the field defined by the SBF function might not actually be from a true source. This problem is tackled by defining a modified likelihood function which accounts for speech activity with use of a uniform distribution $\mathcal{U}$ for non-speech segments:

$$p(\mathtt{Y}|\mathtt{X}) = q_0 \mathcal{U}(l_{\mathtt{X}}) + \gamma(1 - q_0)[P(l_{\mathtt{X}})]^r, \tag{16}$$

where q_0 is the prior probability that an SBF was measured from a noise segment, γ and r are shaping constants for the $P(l)$ distribution ensuring good source tracking performance. The q_0 can be defined using soft-decision output (without thresholding) $\alpha(k) \in [0, 1]$ which is direct complement of the q_0 and therefore the time varying version becomes

$$q_0(k) = 1 - \alpha(k), \tag{17}$$

where implementation of the VAD function $\alpha(k)$ varies by application. In work by Johansson *et al.* [11] it corresponds to the estimated speech signal level, derived from the SNR and noise power estimates of the VAD.

5 REFERENCES

[1] J. Ramírez, J. M. Gorriz, and J. C. Segura, *Robust Speech Recognition and Understanding*, chapter Voice Activity Detection. Fundamentals and Speech Recognition System Robustness (I-Tech Education and Publishing) (2007).

[2] 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 **ASSP-32**, 1109–1121 (1984).

[3] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold", IEEE Transactions on Audio, Speech, and Language Processing **14**, 412–424 (2006).

[4] J. Ramírez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "An effective

subband OSF-based VAD with noise reduction for robust speech recognition", IEEE Transactions on Speech and Audio Processing **13**, 1119–1129 (2005).

[5] J. H. L. Hansen, R. Huang, B. Zhou, M. Seadle, J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "SpeechFind: Advances in spoken document retrieval for a national gallery of the spoken word", IEEE Transactions on Speech and Audio Processing **13**, 712–730 (2005).

[6] U. Yapanel and J. H. L. Hansen, "A new perspective on feature extraction for robust in-vehicle speech recognition", in *Proc. Eurospeech*, 1281–1284 (2003).

[7] S. Dharanipragada and B. Rao, "Mvdr-based feature extraction for robust speech recognition", in *ICASSP* (Salt Lake City, UT) (2001).

[8] R. Huang and J. H. L. Hansen, "Unsupervised audio segmentation and classification for robust spoken document retrieval", in *IEEE ICASSP*, 741–744 (Montreal, QC, Canada) (2004).

[9] K. Pearson, "On lines and planes of closest fit to systems of points in space", Philosophical Magazine **2**, 559–572 (1901).

[10] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking", EURASIP Journal on Advances in Signal Processing **2007** (2007), Article ID 50870, 11 pages.

[11] A. M. Johansson, E. A. Lehmann, and S. Nordholm, "Real-time implementation of a particle filter with integrated voice activity detector for acoustic speaker tracking", in *Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS'06)*, 1004–1007 (Singapore) (2006).

[12] E. A. Lehmann and A. M. Johansson, "Experimental performance assessment of a particle filter with voice activity data fusion for acoustic speaker tracking", in *Proceedings of the IEEE Nordic Signal Processing Symposium (NORSIG'06)*, 126–129 (Reykjavik, Iceland) (2006).

Adaptation Techniques for Noise Robust Automatic Speech Recognition

Tuomo Raitio

Helsinki University of Technology
Department of Signal Processing and Acoustics
email: tuomo.raitio@hut.fi

Abstract

Adaptation techniques can be used to alleviate the problem of mismatch between the training and test data in automatic speech recognition. A variety of adaptation techniques exist. Generally, most of the methods are used for speaker adaptation, while the adaptation of the environment is a less studied issue. In this review, various adaptation methods are described and their applications in noise robust automatic speech recognition are discussed.

1 INTRODUCTION

The performance of an automatic speech recognition (ASR) system is highly dependent on the good match of the training and test data. In clean speech conditions, the performance of a speaker dependent (SD) automatic speech recognition system is generally good, but in real-life conditions with various speakers, the performance may be greatly reduced. In practical applications of ASR systems, there are at least two main issues that cause severe mismatch: (a) the voice of the speaker is different from the voice of the training data, and (b) the environment, exhibiting additive and convolutional noise, clearly differs from the recording environment of the training corpus. The speaker and environment affect greatly the recognition accuracy, since the models of the trained hidden Markov models (HMMs) do not correspond to the parameters of the recorded real-life speech.

Adaptation techniques can be used to alleviate the problem of mismatch between the training and test data. Adaptation is a method, where a small amount of speech data is used to transform the HMM models constructed from the original training data in order to easily build an HMM system that has the properties of the adaptation data. Speaker adaptation is widely studied issue, and speaker adapted systems provide substantial improvements compared to speaker independent (SI) speech recognizers, and even close the performance of speaker dependent (SD) systems trained with full corpus. However, in normal speaker adaptive systems, the environmental mismatch remains a problem. The effect of noise on the recognition accuracy have been usually compensated by using speech enhancement or modifying the feature extraction, but lately, adaptation techniques have also been used for compensating for the environmental mismatch. In this review[1], different adaptation methods are described along with their applications in noise robust automatic speech recognition.

[1]This review is partly based on paper of P. C. Woodland: *Speaker Adaptation for Continuous Density HMMs: A Review* [1].

2 ADAPTATION

HMM-based automatic speech recognition systems use adaptation in order to easily decrease the difference between the training and test data. This is achieved by using a small amount of speech data, representing the speech and environment of the test data, for modifying the original statistical models (HMMs) to resemble the test speaker and conditions.

Adaptation can be operated in a number of modes. It can be either *supervised* or *unsupervised*. In supervised adaptation, the transcription of the speaker specific adaptation data is known. In unsupervised adaptation, the transcription must first be estimated (recognized). Adaptation can also be performed in *static* (block) mode, in which the adaptation is performed once at the beginning, or it can be *dynamic* (incremental), in which the system continues to adapt over time.

The transformation from the original HMMs to adapted HMMs can be performed with various methods. Usually, the transformation is applied only to the mean vectors of the Gaussians in the mixture, but also the covariances or mixture weights can be adapted.

3 ADAPTATION METHODS

Adaptation methods fall into three main families based on (a) maximum a posteriori (MAP) adaptation (b) linear transforms of model parameters and (c) speaker clustering and speaker space methods. The performance of the methods vary in different applications. In the next few sections, various adaptation methods are described along with their strengths and weaknesses.

3.1 Maximum A Posteriori (MAP) Adaptation

Most HMM-based ASR systems are trained using maximum likelihood (ML) estimation; the parameters values, λ, are chosen so that the likelihood of the training data, $p(x|\lambda)$, is maximized. In MAP estimation, the parameters are defined from the posterior distribution $p(x|\lambda)p_0(\lambda)$, where $p_0(\lambda)$ is the prior distribution of the parameters, representing the model before adaptation. In standard MAP approach, for a particular Gaussian mean, the MAP estimate for mean is

$$\hat{\mu} = \frac{\tau\mu_0 + \sum_{t=1}^{T} \gamma(t) o_t}{\tau + \sum_{t=1}^{T} \gamma(t)} \tag{1}$$

where τ is a meta-parameter, defining the bias between the prior mean and the ML estimate of the mean of the adaptation data, $o(t)$ is the adaptation vector at time t from a T length set and $\gamma(t)$ is the probability of this Gaussian at time t. Similar formula can be used to adapt the mean and mixture weights in the system.

One key advantage of MAP approach is that it converges towards ML estimate as the adaptation data increases towards infinity. However, the main drawback of MAP approach is that it only adapts the parameters that are observed in the adaptation data. Therefore, extensions to MAP adaptation have been developed.

In regression based model prediction (RMP) [2] approach, the aim is to find correlation

between the parameters of an HMM system and use these linear regression relationships to update poorly adapted or unobserved parameters based on the well-adapted parameters. In this approach, first, correlation between the mean values of the Gaussians with the speaker-dependent changes is estimated. After the estimation, standard MAP is used for adapting the models, and the well-adapted data is then used for generating parameter estimates for the poorly adapted data using the correlation. The final mean value is a linear combination of the initial MAP estimate and the predicted value. In general, RMP converges to the same word error rate (WER) than MAP, but outperforms MAP for small amount of adaptation data.

Structural MAP (SMAP) adaptation [3] is another technique that tries to enhance the adaptation of unobserved data. In SMAP adaptation, the Gaussians in the system are organized into a tree structure according to acoustic similarity, and a mean offset and a diagonal variance scaling term are recursively computed for each layer, starting from the root node. At each level in the tree, the distribution from the node above is used as a prior. The tree structure ensures that unobserved data are adapted in the same way with similar speech sounds.

3.2 Linear Transformation Adaptation

Another approach to the speaker adaptation problem is to estimate a linear transformation of the model parameters to construct a better model. This approach provides fairly rapid and robust adaptation.

There are various schemes that use linear transformations. In maximum likelihood linear regression (MLLR) [4], the Gaussian mean parameters are updated according to

$$\hat{\mu} = \mathbf{A}\mu + b \tag{2}$$

where $\mathbf{A}$ is an $n \times n$ matrix and b is an n dimensional vector. The parameters $\mathbf{A}$ and b are estimated so that the likelihood of the adaptation data is maximized. Expectation-Maximization (EM) algorithm is used for searching the optimal parameters.

However, there is a tradeoff between *global* transform and using more precise transforms that apply to smaller number of parameters (*local*). For example, a general transform for all the Gaussians may not be accurate, since the differences in data might be phone-specific. One solution to this problem is to use Regression Class Tree [5]. The idea is similar to SMAP; the Gaussians are arranged so that they are close in the acoustic space, and the same clusters undergo the same transformation.

In constrained MLLR (CMLLR) [6], the means and variances of Gaussians are transformed simultaneously according to

$$\hat{\mu} = \mathbf{A}_c\mu - b_c \tag{3}$$

$$\hat{\Sigma} = \mathbf{A}_c^T\Sigma\mathbf{A}_c \tag{4}$$

This is equivalent to transforming the observation vectors such that

$$\hat{o}_t = \mathbf{A}_c^{-1} o_t + \mathbf{A}_c^{-1} b_c \tag{5}$$

This enables the direct transformation of the observation vectors of the incoming data, which can be convenient in some applications.

One disadvantage of MLLR is the possibility of over-training on the adaptation data. In practice, limits can be set to preserve the performance better than the original speaker independent (SI) system, but methods that do not require this would be preferable. One solution is to use MAP-like interpolation between the original mean and the MLLR estimated mean [7]. In these methods, prior distributions are defined for the transformation matrix parameters, yielding MAPLR adaptation. This can also be extended to structural maximum a posteriori (SMAP) adaptation, yielding SMAPLR.

3.3 Speaker Clustering and Speaker Space Adaptation

Previous methods have not explicitly utilized information about the characteristics of an HMM set for particular speakers. However, forming groups of speakers and utilizing the information from a specific group can be very useful in speaker adaptation. Recently there has been interest in cluster adaptive training [8] and eigenvoice techniques [9], which both form a weighted sum of speaker HMMs, and use this to represent specific speaker.

Cluster adaptive training (CAT) represents a speaker as a weighted sum of individual speaker cluster models. The models have common variance and mixture weights, and only the Gaussian mean values vary. Thus, a mean of particular speaker is represented as

$$\hat{\mu} = \sum_c \lambda_c \mu_c \tag{6}$$

where the parameters of the model are the speaker-specific λ_c, which define the cluster weights and μ_c is the corresponding mean of Gaussian in cluster c. Thus, for a particular set of canonical speaker cluster models and some adaptation data, maximum likelihood weights can be derived.

Eigenvoice technique also performs speaker adaptation by forming models as a weighted sum of canonical speaker HMMs and adapts only the mean vectors, but the canonical speakers (eigenvoices) are found using principal component analysis (PCA). The eigenvoices of with largest eigenvalues $\bar{\mu}_e$, $e = 1, ..., E$ are chosen as an E-dimensional basis set. The parameter vector of the adaptation data for particular speaker μ can be represented as an eigendecomposition

$$\mu = \sum_{e=1}^{E} w_e \bar{\mu}_e = Mw \tag{7}$$

where $w = [w_1, ..., w_E]^T$ are the eigenvalues that represent the speaker, and $M = [\bar{\mu}_1^T, ..., \bar{\mu}_E^T]^T$ is the eigenspace. Using EM algorithm, the maximum likelihood weights of the eigenvoices $\hat{w}$ are computed for the particular speaker (observation o) by

$$\hat{w} = \arg\max_w \log L(o|\mu = Mw). \tag{8}$$

CAT and eigenvoice technique yield good performance for a very small amount of adaptation data, but the problem is that the performance does not converge to speaker dependent system. One solution to this is to use prior distribution for MAP adaptation based on eigenvoices [10], which yields both rapid adaptation and convergence to the MAP solution.

4 ADAPTATION FOR NOISE ROBUST ASR

In general, the adaptation methods described above all modify the acoustic models to better match with some adaptation data. Although adaptation methods are usually utilized for decreasing the mismatch due to the speaker, the methods can be used for decreasing the mismatch from the environment as well. Since general MAP and MLLR type techniques do not include any kind of in-built speaker model, these techniques can be directly used for cases of acoustic environment mismatch, or a combination of speaker and environment mismatch. However, in an ideal case the environmental adaptation and speaker adaptation should be different processes. This would be useful in order to build a speaker adapted system that was independent of the acoustic environment, providing fast and easy speaker adaptation in real-life conditions.

In [11] it was suggested to use the eigenvoice method to model speaker changes and then use MLLR adaptation to compensate for the environmental mismatch. The proposed method comprises the following steps. First, an eigenspace is trained on a large clean speech database D_0, and a small amount of data D_1 is recorded in real-life conditions. The speech to be recognized in the same real-life conditions is denoted by D_2. Second, for each speaker in D_1, maximum-likelihood eigenspace decomposition (MLED) is computed, which projects the speakers in the reference space. Then, MLLR is used to compute the transformation between the noisy and clean data for all speakers, making the transformation to focus on environmental variation only. This step yields environment-dependent speaker independent (SI) system. Finally, in real-life speech recognition, for each speaker in D_2, MLED is computed in the reference space in order to adapt with the speaker, and then the estimate is rescaled with the effect of the environment using MLLR. This yields both speaker and environment adapted ASR system.

In another approach, acoustic factorization is proposed [12], where each unwanted factor affecting the acoustic signal is separately modeled, which allows the appropriate factors to be factored-in for a particular target domain. In the proposed method, eigenspace/CAT was used to project the transformation between noisy and clean speech data, and MLLR was used to perform the speaker adaptation. The acoustic factorization approach may lead to rapid combined speaker and environment adaptation, and allow the speaker transform parameters to be retained across noise conditions.

5 SUMMARY

Adaptation is an effective method for decreasing the mismatch between training and test data in automatic speech recognition. There is a variety of different adaptation methods, which fall into three main categories. MAP-based adaptation has the advantage that it

converges to ML estimate as the adaptation data increases. However, the major drawback of MAP-adaptation is the speed of adaptation in large HMM systems. This can be alleviated by using e.g. regression based model prediction (RMP) or structural MAP (SMAP). The transformation based approaches are effective with very small amounts of adaptation data, but over-training is a problem. This can be prevented by using combined adaptation with MLLR and MAP or SMAP. Speaker clustering and speaker space methods form models as a weighted sum of speaker HMMs. In cluster adaptive training (CAT), the model is formed as a weighted sum of individual speaker cluster models, whereas in eigenvoice based method, PCA is used to find a set of supervectors that define the speech space, and a speaker model is formed as a weighted sum of these supervectors. CAT and eigenvoice techniques are efficient with a very small amount of data, but do not converge to the MLLR or MAP solutions.

In general, all of these adaptation methods can be used for decreasing the mismatch from speaker, environment, or both. However, in an ideal case, the environmental adaptation and speaker adaptation should be different processes. This would be useful in constructing speaker adapted system that was independent of the acoustic environment, enabling fast speaker adaptation in new real environments. A few such methods have been already described, using CAT/eigenvoices and MLLR for speaker and environment adaptation (and vice versa). However, there is still not much research on environment adaptation.

6 REFERENCES

[1] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review", in *Proc. ISCA Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, 11–19 (2001).

[2] S. Ahadi and P. Woodland, "Combined bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models", Computer Speech & Language **11**, 187–206 (1997).

[3] K. Shinoda and C. H. Lee, "Structural MAP speaker adaptation using hierarchical priors", in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 381–388 (1997).

[4] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech & Language **9**, 171–185 (1995).

[5] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression", in *Proc. ARPA Spoken Language Technology Workshop*, 104–109 (1995).

[6] V. Digilakis, D. Ritchev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures", IEEE Trans. SAP **3**, 357–366 (1995).

[7] S. Goronzy and R. Kompe, "A MAP-like weighting scheme for MLLR speaker adaptation", in *Proc. Eurospeech'99*, 5–8 (1999).

[8] M. J. F. Gales, "Cluster adaptive training of hidden Markov models", IEEE Trans. SAP **8**, 417–428 (2000).

[9] R. Kuhn, J. C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space", IEEE Trans. SAP **8**, 695–707 (2000).

[10] H. Botterweck, "Anisotropic MAP defined by eigenvoices for large vocabulary recognition", in *Proc. ICASSP'01*, 353–356 (2001).

[11] P. Nguyen, C. Wellekens, and J. C. Junqua, "Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments", in *Eurospeech'99*, 2519–2522 (1999).

[12] M. J. F. Gales, "Acoustic factorisation: Theory and initial evaluation", Technical Report, CUED/FINFENG/TR.419, Cambridge University Engineering Dept. (2001).

Machine vs. Human Speech Recognition

Prem Raj Adhikari
Department of Information and Computer Science
Helsinki University of Technology
prem.adhikari@tkk.fi

Abstract

Automatic Speech Recognition (ASR) has always been a difficult and interdisciplinary subject, involving at least Computer Science (algorithms, pattern recognition), Electrical Engineering (speech modeling, signal processing), Linguistics (phonetics, phonology, speech structure) and Physics (acoustics). Today, with the advancement of technology such as the increase in computational power, storage capacity etc and more importantly the increase in research in the field of ASR we are close to replicating human performance in Speech Recognition. However, it has to be noted that even with the current state-of-the-art ASR systems; general belief is that the human are the better recognizers of speech than the machines. Both human and machines have some advantages and disadvantages of their own when it comes to speech recognition. For example, human are more robust and efficient but they are also lazy, get tired or bored. Humans are also only good recognizers for few specific languages. So it is not surprising that that we already have few ASR systems that perform better than human for some special cases. This paper discusses the reasons for human performing better than the machines. Comparison between human and machine speech recognition is carried out. This paper also discusses the how human perception of the speech can be exploited to improve the performance of ASR systems as well as other techniques that could be useful to replicate human performance in ASR systems.

1 INTRODUCTION

In recent years, technological development has left a positive impact on the speech recognition technologies. There has been significant advancement in computing power of the hardware which allows real-time solution to the computationally complex speech recognition algorithms based on machine learning aspects and involving complex mathematical and statistical calculations. Increase in storage capacity has allowed the storage and retrieval of huge amount of speech data and corpora with large vocabulary sizes. With ever growing research in the field of speech recognition the sizes of speech corpora has also dramatically increased. Increasing research also means that several computationally fast algorithms have been developed.

Large vocabulary talker independent recognizers provide error rates that are less than 10% for reading sentences recorded in quiet environment [1]. However, these

advances are not enough for real time commercial speech recognition system. Although few simple commercial applications have been developed, for example digit recognition system (vocabulary size only 10), there is no denying the fact that ASR systems can be widely deployed only when its performance is closer to the performance of the human. Significant gap exists between human and machine in terms of speech recognition capabilities. This gap is because of some of the inherent capabilities of the human and various problems still confronting the speech recognition society at present.

Studying human speech recognition is a difficult task. The observations are not direct. Components of the human speech recognition system cannot be separated and tested independently. Although many studies have been carried out, working of the brain which plays an important in human speech recognition system is a mystery not yet fully solved. Experiment on human speech recognition (HSR) is very expensive and difficult to carry out [1].

The aim of this paper is to present the current situation in the research of the speech recognition system, compare it with the human speech recognition system. Performance of the current state of the art ASR system is also evaluated. The aim is not only to compare machine speech recognition system with human speech recognition system but also to present some ideas how study of human speech recognition can be useful for implementation in machine.

2 WHY HUMANS ARE BETTER AT SPEECH RECOGNITION?

Human speech recognition system is built by nature. With centuries of evolution and improvement, long hours of training human speech recognition system are extremely efficient, robust to noise (or signal distortions) and flexible. As a general belief it is difficult (sometimes even impossible) to beat nature. Presence of various natural features, some implementable in machines and some usually not, make humans superior in speech recognition. There is no motivation in plain duplication human speech recognition system. This is not possible and would never be useful. Human still has some problems in speech recognition and machines have some advantage in some regards. So, we can study few specific features of human speech recognition and human perception of speech which could be useful for future research in speech recognition technology.

2.1 Adaptation and Normalization:

Human adapts to difference in speakers, speaking styles and pace of speech (eg. Finnish speaking slow English and Australian speaking fast English hardly seems to be a problem for human). Even if ASR systems are used to handle these differences, they require significant amount of training data and time to adapt. On the contrary, if the

ASR systems are not adapted to the new environment, the results can be poor. In ASR, new models developed to adapt to a changing environment (noise, speaker, speaking styles, pace of speech etc) should not require new training but simple changes in the parameter of the models.

2.2 Prediction Capability and OOV handling

Number of experiments have been carried out to prove that human are better than machines (for example: statistical methods such n-grams) to estimate the upcoming word in the speech stream. Although the perplexity of the human can be poor compared with the machines, machines find it difficult to process from left to right.

Errors can occur when the given input is not in the vocabulary, pushing the system to erroneously recognize the word. This problem is often referred as an Out of Vocabulary (OOV) issue. The source of OOV words are linked to limited size of the training data and to increasing vocabulary in certain domains, e.g.; scientific terms or slang. Humans are better at handling OOV words.

2.3 Word perception models:

Various researchers from various fields have suggested numerous speech perception and word recognition models such as analysis by synthesis (Stevans, 1960), first order context sensitive coding(ERIS)(Marcus, 1981). However, several experiments have proved that HMM (Hidden Markov Model) based word recognition model is efficient. However, incorporating specific information about the speech that can be easily extracted (eg. Environmental characteristics, speaking style, characteristics related to the speaker, the local speaking rate and word characteristics such as word stress, reduction and co-articulation) will permit quick adaptation of model parameters [2].

2.4 Scale Spacing:

Scale spacing in human speech recognition is a process of defining the level of precision for speech understanding either global or detailed. Human are very good at scaling. They can focus their attention exactly according to their current requirement of speech perception. If the requirement is detailed (for example: direction to some place) we can easily focus our attention and remember the speech for a long time. If the details aren't necessary we can just take in the required knowledge and forget about the nitty-gritty details. However, it is difficult for the machine to adapt to such scale spacing. The precision of knowledge of speech recognition is one level only.

2.5 Pronunciation Variations:

People pronounce words differently and pronunciation of same word differs with respect to the context. For example, pronunciation of words such as character, pi (π) is different in different places. However, ASR systems are often implemented as speaker independent and simply take into consideration only the standard pronunciation without considering the context and the speaker.

Human on the other hand have a good knowledge about word boundary, syllabus etc. which makes it easier for the human to understand the speech.

2.6 Noise and Reverberation Robust:

Experiments carried out by Lippmann [1] have shown that human performance is not affected by noise and reverberation when the level of noise is above 0 dB. Only two aspects that affect the human performance are native language of both the speaker as well as the speaker and the size of the vocabulary.

2.7 Robust to spectral and temporal distortions:

ASR systems are trained on phoneme, morpheme, tri-phone word templates. Now when ASR has to recognize a word, it selects the word with shortest distance or greatest similarity to the word in its vocabulary. So when the signals are distorted measures of similarity and shortest distance will also be effected thus effecting the result of recognition. Similar to machines, humans also learn from examples. However, the experiments (Grossberg, 1986) have found that human are more flexible and adaptable.

2.8 Additional Contextual Information:

Speech is an acoustic signal. However, when people communicate various features such as facial expression, movement of the hands, blinking of the eye etc. give additional information about the speech. “Information is in the audio, video is for entertainment” R. Schwartz. However, video plays an important role in human speech perception as well. Human use these additional information efficiently in speech recognition to understand the speaker. Popularity of audio visual synthesis, automatic generation of voice and facial animation from text (also known as multimodal speech synthesis), is increasing day by day.

Human auditory system divides the acoustic signal into frequency bands, called critical bands, which are used as independent inputs to human recognition mechanisms. In experiments carried out by Fletcher et. al. [10] it was discovered that biological combination scheme obeys the following regularity.

$$Q_X(X) = 1 - \prod_{\forall i} \left(1 - Q_{X,i}(X)\right)$$

Where X is the input from several non-overlapping bands centered around frequency f_i. The $Q_X(X)$ is the minimal context. This means that error is committed only if none of the channels yields correct classification. Human always take into account the contextual information along with acoustic information. Often contextual constraints can be implemented in ASR as the probability of various utterances thus limiting the number of possible words.

2.9 Auditory Modeling:

Neuro-mechanical signal processing in the human peripheral auditory system is very complex. It has functionalities that we need not consider for speech recognition. So, instead of plain duplication of human system, the speech recognition methods using HMM and other techniques already have a better performance. However, careful feature extraction can improve the performance of the machine recognizers.

2.10 Prosody-driven recognition:

The acoustic parameters responsible for prosody are generally considered to be fundamental frequency, duration, energy, and spectral slope as a function of time [2]. Prosody is important in human speech recognition for determining the word boundaries etc. Stress and syllables in words if detected as the humans do, the performance of ASR systems are bound to improve.

Some of the consistent speech characteristics such as pitch information, durational variability, spectral reduction and co-articulation, quick adaptation to speaker, style and communication channel, communicative expectation, multi-modality, binaural hearing etc are often rejected in speech recognition [2]. If we could use these features in ASR, the performance should improve.

3 SPEECH SYNTHESIS (MACHINE GENERATED SPEECH)

Machines have already been used to produce speech for various purposes. Small vocabulary systems such as flight announcement systems, speaking clock etc produce sound close to that produced by human. Unlimited vocabulary speech synthesis is also possible using rules but with increase in vocabulary there will be decrease in quality and interchangeability. Current hardware and storage systems are not yet capable of handling such large amount of data and rules for real time speech production.

Performance of speech synthesis systems also depends on the context. For example, the read aloud Wall Street Journal (WSJ), later extended to many more

newspapers in the North American Business (NAB) news has good performance with error rate of about 10%. However, producing speech synthesizer for uttering the names and addresses of people is difficult because context in the case of WSJ is simple but context in case of names and addresses is complex with each names and addresses being different.

4 COMPARISONS HUMAN AND MACHINE SPEECH RECOGNITION

Lippmann [1] pioneered the comparison of human and machine speech recognition techniques.

4.1 The Comparison

Experiments were carried out on six different evaluation corpora: Digit and digit strings, 26 alphabet letters, North American Business News, Resource Management, Switchboard Spontaneous Conversation (Telephone) and Switchboard word spotting (Telephone). The four basic objectives of carrying out the comparisons were: to focus the research activities to narrow the gap between human speech recognition and machine speech recognition, to study the human behavior with regard to speech recognition like how human adapt to changes and also bridge the gap between human and machine speech recognition research.

Lippmann [1] used six different evaluation corpora used which were designed for speaker independent speech recognizers. The corpora were designed for different applications of speech recognition technology. Word error rate constitutes substitutions, insertions, deletions, and word insertions. Small numbers of OOV words are included as test words.

All the speech corpora focus on dictation task where the objective is to identify the spoken word except for the switchboard corpus which was used both for dictation and word spotting i.e. to detect 20 common words in conventional telephone speech.

Table below gives the characteristics of the six evaluation corpora used for comparison of human and machine speech recognition.

Table 1: Characteristics of 6 speaker independent speech recognition corpora by Lippmann [1].

Corpus	Description	Number of Talkers	Vocabulary Size	Number of Utterances	Total duration	Recognition Perplexity
TI Digits	Read digits	326	10	25,102	4 hrs	10
Alphabet Letters	Read alphabets	150	26	7,800	1 hr	26
Resource Management	Read sentences	109	1000	4,000	4 hrs	60-1000
North American Business News(NAB)	Read sentences	84-284	5000-unlimited	7,200-37,200	12-62 hrs	45-160
Switchboard continuous speech recognition	Spontaneous telephone conversations	70-543	2000-unlimited	35-2400 conversations	2-240 hrs	80-150
Switchboard word spotting	Spontaneous telephone conversations	70	20 keywords	2000 keyword occurances	2 hrs	-

Experimental Design:

Before the experiments were carried out transcription of the corpus is carried out by highly motivated listeners. The metric used for the comparison is word error rate. Other difficult metrics such as recognition time, training time, vocabulary size, type of noise etc were not considered. Same corpus was used for both human and machine except for the resource management corpus with a null grammar and for the Alphabet letter corpus. Experiments use the results from six different ASR systems that perform best on specific corpora.

For humans also committee majority vote among multiple humans was used instead of average error rate. Spellings were also corrected to eliminate OOV responses. They were motivated for better performance as they would receive extra pay if they made fewer errors.

Table 2 below shows the comparison of the best performing ASR and majority of the committee vote for the speech corpora given in the table 1 above.

Table 2: Summary of the error rates for Human and best performing ASR by Lippmann[1].

Corpus	Machine Percentage WER	Human Percentage WER
TI Digits	0.72	0.009
Alphabets	5	1.6
Resource Management	17	2
North American Business News	6.6	0.4
Switchboard continuous speech recognition	43	4
Switchboard word spotting	31.1	7.4

For example in the case of TI digits corpora, 25 listeners typed their responses in keyboard. The average digit string or pre-utterance rate for vocoded speech was 0.105%. It further reduces to 0.01% when the majority vote from a committee of three listeners was used. When wide band speech was used the string error rate for the individual listeners dropped to 0.009%. On the other hand machine error rate is 0.72%.

4.2 Benchmarking Human performance for continuous speech recognition

Deshmukh, N et. al. [8,9] used LVCSR and low SNRs to compare human and machine speech recognition. They use two evaluation corpora: ARPA's CSR'94 Spoke 10 and CSR'95 Hub 3. This study was primarily carried out to set a benchmark for machine error rates. The aim was also to study the effect of various factors such as amount of noise, speaker characteristics, artifacts introduced by the placement and type of microphone, properties of spoken material on the human speech recognition.

Experimental Design:
Both native and non-native speakers of English were used as subjects. Test data was systematically distributed so that ultimate human performance could be closely approximated. Different microphone conditions were also used to study the effects of type of microphone on the speech recognition.

Evaluation Corpora: Two evaluation corpora used had the following features.

94 Spoke 10: The transcripts of 113 sentences were recorded from 10 speakers with about 11 sentences per speaker. SNR of levels (22dB, 16dB and 10dB) were added to create a total of 452 utterances.

95 Hub 3: The corpus of 20 paragraphs with 15 sentences in each paragraph which were drawn from broadcast news. Each utterance was recorded using two microphones and hence there were 600 utterances.

Listeners were not motivated as they were not given any extra benefits for correct answers. Each listener listened to a different speaker with different condition each time. Committee transcription was used from the transcription by a group of 3 listeners for each utterance under each condition. Spelling errors and other attention span errors were corrected as a post processing step.
The results of the experiment are described in the table below.

Table 3: Evalation of Human Recognition Performance on 94 Spoke 10 Corpus on noise condition. Adapted from Neeraj Deshmukh et. al. [8]

Vocabulary Evaluation	**Listener**	**Percentage WER on Noise Condition(94 Spoke 10 Corpus)**				
		None	**22dB**	**16dB**	**10dB**	**All**
Open Vocabulary	Overall	2.0	2.1	2.1	2.1	2.1
	Committee	1.0	1.4	1.2	1.2	1.2
Augmented Vocabulary	Overall	1.0	0.8	0.8	1.1	0.9
	Committee	0.6	0.3	0.4	0.7	0.5

Table 4: Evalution of Human Recognition Performance on 95-Hub 3 Corpus on Microphone Condition by Deshmukh et al. [8,9].

Vocabulary Evaluation	**Listener**	**Percentage WER on Microphone Condition(95 Hub 3 Corpus)**				
		Mic s	**Mic b**	**Mic f**	**Mic g**	**All**
Open Vocabulary	Overall	2.0	1.9	1.9	3.3	2.2
	Committee	0.8	1.0	0.2	1.5	2.2
Augmented Vocabulary	Overall	1.5	1.4	1.6	2.2	1.6
	Committee	0.3	0.6	0.1	0.8	0.4
Best Performing ASR(Machine)		6.6	8.1	10.3	23.9	10.0

Mic s: Sennheiser HMD-410 close-talking;
Mic b: Shure SM-58 boom-mounted;
Mic f: Audio Technica AT851a Micro Cardioid Condenser
Mic g: Radio Shack 33-1060 Omni Electret

As can be seen from above that human error rate are statistically insignificant as it is less than 1%. It can also been seen that error rates are similar for all the noise levels.

Hence the effect of noise on human recognition is negligible. Experiments also found that most of the error was on the recognizing the proper nouns.

5 WHY MACHINES CAN BE BETTER?

Comparison with human shows that ASR systems have severe pitfalls in speech recognition. It has to be mentioned that some modern and advanced ASR systems have excellent performance in noise free condition. However, the performance degrades drastically when the noise and other anomalies degrade the quality of speech signal. For machines error rates increase roughly to 40% for spontaneous speech and to 23% with channel variability and noise [1]. On the other hand human error never exceeds 5% in quiet as well as degraded conditions. Therefore, it is evident that human speech recognition thrives in noise and difficult task such as continuous speech.

Despite several superiorities of human speech recognition system, there are certain aspects where machine has advantage over human.

5.1 Inherent Advantages of Machines over humans

Human can get tired easily, can't be asked to work more than 8 hours a day. Human can also be distracted by other interesting stimuli (e.g. visual stimuli) and they tend to get bored working on the similar tasks time and again. They can get uninterested on a topic and subject. The human nature the feeling of emotions (happiness and sadness) also plays an important role in degrading human performance. Another significant flaw in human can be of language. Normally, human are better at listening of few languages (mostly 1 or 2, in almost all the cases less than 5). It would be difficult and would take significant time to train and adapt a person who is good listener in English to listen with same performance in Finnish. On the other hand, with some simple and minor changes in the model parameters, speech corpora and with few hours of training machine can adapt and replicate its performance in an entirely new language.

5.2 Advancement

Human evolution is pretty slow process. It takes centuries for humans to develop something new. For example we have not been able to get rid of the vestibule organs for such a long period of time. On the other hand, the development of the ASR methods is dramatic as has been shown in the last decade. Numerous researches are being carried out in various fields of ASR in several leading universities and laboratories around the world. If the rate of advancement carries on at the current rate than it will not be long before we can see ASR systems that outperform human in all the circumstances.

6 HOW CAN MACHINES BE IMPROVED?

6.1 Use of glimpsing model

Glimpses are the regions in the spectro-temporal regions in which target signal is least effected by background [3]. Humans recognize speech in glimpses of speech (reliable evidence) and treat other spectro-temporal parts as missing evidence. It is similar to computer vision. If a car is seen passing by, which is partially covered by some object, mind can fill things that can't be seen at the moment thus realizing the car as a whole object based on the glimpse. Experiments carried out by Cooke [3] found that information in glimpses could be useful for consonant identification and suitable to be used for computation as well.

6.2 Bridging Communities of ASR and HSR

Although the aim of both ASR and HSR is to extract linguistic information from acoustic signal, there exist two distinctly separate communities for research on the respective fields. In spite of some differences in their objectives and features, there is a overwhelming possibility of establishing contact points between the two research communities. Sadly, there hasn't been much effort from both the research communities to consider bridging the gap between them. Both research communities can reap the benefits if the research communities can establish more contacts between themselves.

6.3 Human Factor Cepstral Coefficients (HFCC)

HMM and MFCC has some severe limitations. First is foremost is HMMs neither claim nor model biology. Secondly, HMMs are piecewise-stationary, while speech is continuous and nonstationary. Thirdly, HMM assumes that frames of speech are independent and identically distributed. Finally, state pdf(probability distribution function) are data driven. So, these limitations mean we must come up with a new model. Psychoacoustic experiments have found several properties human auditory system that could be used in feature extraction algorithm. Mel frequency scale, band pass sensitivity, and temporal and instantaneous masking have already been implemented in current ASR system. HFCC has been purposed as a new model to overcome the above problems in HMM and MFCC. HFCC includes critical bandwidth information into the filter bank design of MFCC. Independent optimization of filter bandwidth possible through HFCC is even better than the biologically inspired. HFCC-E (HFCC with linear E-factor) is an advancement of HFCC where the ERB (Equivalent rectangular bandwidth) used is linearly scaled.

6.4 Energy Redistribution algorithms

Speech signal in the time domain contains several features like gender, age, emotional state of the speaker. Thus speech stream has variable concentration of speech over time. Energy redistribution algorithms conserve signal energy and preserve naturalness while intelligibility of the linguistic information is increased (as determined through human listening tests)[7] exploiting the variable concentration of information over time.

6.5 Creation of Novel Classifier

Recent researches have been successful in studying sensory processing in biology. New theories on information processing are being formulated based on the chaotic properties observed in the biological systems. Already some systems such as Freeman's KIII network has been developed which is yet not a replacement for the HMMs although it better models the non-convergent signals detected in biology.

6 CONCLUSIONS

In spite of the recent developments in the various fields of speech recognition, the current ASR systems fall short of human in performance. Relentless research in the field of speech recognition should in the near future produce a generic speech recognition system that can perform at the capacity of human beings. Studying the details of human speech recognition could be beneficial for developing ASR that could match the performance of human recognition system. Some features of human speech recognition system such as contextual information and lower level phonetic modeling could be used to enhance the performance of ASR. There is a need to bridge the gap between research in ASR and HSR community.

7 REFERENCES

[1] Lippmann R. P. (1997) Speech recognition by machines and humans, Speech Comm. 22, 1-15.

[2] Luis C. W. Pols(1999) Flexible, Robust, and Efficient Human Speech Recognition. Institute of Phonetic sciences, University of Amsterdam, Proceedings of ICPhS

[3] Cooke M. P. (2006) A glimpsing model of speech perception in noise, J. Acoust. Soc Am. 119(3), 1562-1573

[4] Misha Pavel, Malcom Slaney and Hynek Hermansky(2009) Reconciliation of Human and Machine Speech Recognition Performance, International Conference on Acoustics, Speech and Signal Processing, IEEE

[5] Jason J. Sroka, Louis D. Braida(2005) Human and Machine Consonant recognition, Speech Communication, Elsevier

[6] Cooke M. P. and Ellis D. P. W. (2001) The auditory organization of speech and other sources in listeners and computational models, Speech Comm. 35, 141-177

[7] Mark D. Skowronski(2004) Biologically Inspired Noise-Robust Speech Recognition For Both Man And Machine, PhD Dissertation

[8] Deshmukh, N. Duncan, R.J. Ganapathiraju, A. Picone, J. (1996) Benchmarking human performance for continuous speech recognition Inst. for Signal & Inf. Process., Mississippi State Univ., MS

[9] N. Deshmukh, A. Ganapathiraju, R.J. Duncan and J. Picone (1996) "Human Speech Recognition Performance on the 1995 CSR Hub-3 Corpus", in *Proceedings of the ARPA Speech Recognition Workshop*, Harriman, NY.

[10] Allen J. B., Fletcher H. (1995), The ASA Edition of Speech and Hearing in Communication, published for the Acoustical Society of America by the American Institute of Physics

TKK REPORTS IN INFORMATION AND COMPUTER SCIENCE

TKK-ICS-R9 Elia Liitiäinen, Francesco Corona, Amaury Lendasse
A Boundary Corrected Expansion of the Moments of Nearest Neighbor Distributions. October 2008.

TKK-ICS-R10 He Zhang, Markus Koskela, Jorma Laaksonen
Report on forms of enriched relevance feedback. November 2008.

TKK-ICS-R11 Ville Viitaniemi, Jorma Laaksonen
Evaluation of pointer click relevance feedback in PicSOM. November 2008.

TKK-ICS-R12 Markus Koskela, Jorma Laaksonen
Specification of information interfaces in PinView. November 2008.

TKK-ICS-R13 Jorma Laaksonen
Definition of enriched relevance feedback in PicSOM. November 2008.

TKK-ICS-R14 Jori Dubrovin
Checking Bounded Reachability in Asynchronous Systems by Symbolic Event Tracing. April 2009.

TKK-ICS-R15 Eerika Savia, Kai Puolamäki, Samuel Kaski
On Two-Way Grouping by One-Way Topic Models. May 2009.

TKK-ICS-R16 Antti E. J. Hyvärinen
Approaches to Grid-Based SAT Solving. June 2009.

TKK-ICS-R17 Tuomas Launiainen
Model checking PSL safety properties. August 2009.

TKK-ICS-R18 Roland Kindermann
Testing a Java Card applet using the LIME Interface Test Bench: A case study. September 2009.

ISBN 978-952-248-087-3 (Print)
ISBN 978-952-248-088-0 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)