Uncertainty Decoding on Frequency Filtered Parameters for Robust ASR

Jesús Vicente-Peña, Fernando Díaz-de-María

Department of Signal Processing and Communications
EPS-Universidad Carlos III de Madrid
Avda. de la Universidad, 30, 28911-Leganés (Madrid), Spain
Phone: +34 91 624 9170
Fax: +34 91 624 8749

Abstract

The use of feature enhancement techniques to obtain estimates of the clean parameters is a common approach for robust automatic speech recognition (ASR). However, the decoding algorithm typically ignores how accurate these estimates are. Uncertainty decoding methods incorporate this type of information. In this paper, we develop a formulation of the uncertainty decoding paradigm for Frequency Filtered (FF) parameters using spectral subtraction as a feature enhancement method. Additionally, we show that the uncertainty decoding method for FF parameters admits a simple interpretation as a spectral weighting method that assigns more importance to the most reliable spectral components.

Furthermore, we suggest combining this method with SSBD-HMM (Spectral Subtraction and Bounded Distance HMM), one recently proposed technique that is able to compensate for the effects of features that are highly contaminated (outliers). This combination pursues two objectives: to improve the results achieved by uncertainty decoding methods and to determine which part of the improvements is due to compensating for the effects of outliers and which part is due to compensating for other less deteriorated features.

Key words: Robust speech recognition, spectral subtraction, uncertainty decoding, frequency filtered, bounded distance HMM, SSBD-HMM

Email address: jvicente,fdiaz@tsc.uc3m.es.

1 Introduction

State-of-the-art Automatic Speech Recognition (ASR) systems lose effectiveness due to the *mismatch problem*, i.e., when they operate in environments different from the ones considered in the training stage. One of the causes of the mismatch is the presence of additive noise. This problem has received a lot of attention and has been the subject of numerous research studies (a classic review can be found in Gong (1995)). Feature enhancement is one of the approaches to tackle this problem and consists of compensating for the features before entering the recognition (or decoding) stage, which proceeds as usual. In this work, we study uncertainty decoding methods, which modify the decoding process to incorporate the information about the quality of the feature estimates.

In this paper, we focus on ASR systems that use Frequency Filtered (FF) parameters (Nadeu et al. (1995, 2001); Paliwal (1999)). This parameterization performs as well as the parameterizations in the cepstral domain such as the Mel-frequency cepstral coefficients (MFCC) and has the additional advantage of staying in the log-frequency domain. As we show in the paper, this characteristic allows us to make an easy interpretation of the proposed methods. Furthermore, unlike the MFCCs, FF parameters do not spread a frequency-localized distortion over the remaining frequency bands and, therefore, the effect of the distortion over the entire feature vector is minimized (de Veth et al. (2001b)). Additionally, the compensation of outliers in systems that are designed using the FF parameters is more effective than in those systems using MFCCs (de Veth et al. (2001b,a)).

A previous work, where we also used the FF parameters, showed that the combination of Spectral Subtraction (SS) (e.g. Boll (1979)) and bounded distance HMM (BD-HMM) (Vicente-Peña et al. (2010)), a method inspired by that proposed in de Veth et al. (2001b,a) which mitigates the effect of the outliers in the recognizer, leads to notable improvements of the recognition system performance in the presence of additive noise. In this work, we show that this combination, called SSBD-HMM, can also be effectively complemented with uncertainty decoding. Furthermore, the analysis of this new combination of methods allows us to determine which fraction of the improvements in uncertainty decoding methods is due to the compensation for outliers and which fraction is due to the compensation for other features that are not so highly contaminated.

The proposed method is assessed for the well-known RM1 and Aurora-4 ASR tasks, and our experimental results prove that incorporating the uncertainty of the observations in the decoding process significantly improves the recognition rates.

The rest of the paper is organized as follows. Section 2 introduces uncertainty decoding methods. Next, Section 3 describes our proposal based on the application of uncertainty decoding to the Frequency Filtered parameters. Section 4 briefly reviews the SSBD-HMM method that is used in combination with uncertainty decoding. Section 5 describes the experimental setup and shows the results achieved by the proposed method. Finally, Section 6 gives some conclusions and closes the work.

2 Uncertainty decoding

Feature enhancement methods aim at estimating clean versions of the noisy speech parameters. However, after this estimation, a certain level of uncertainty about the hidden clean parameters still remains, and it is convenient to take into account this information when the decoding process is tackled (e.g. Yoma et al. (1998)). As we briefly review in this section, uncertainty decoding methods also take this information into account during the decoding process. We start reviewing the decoding process in a conventional recognizer and then we review the same process for a recognizer based on uncertainty decoding.

An HMM-based speech recognizer decides on the acoustic unit that has been uttered by computing the maximum likelihood among all the possible acoustic models (see Young et al. (2002) or Rabiner (1989) for details):

$$\lambda = \arg\max_i p(\lambda_i|\mathbf{O}) = \arg\max_i p(\lambda_i)p(\mathbf{O}|\lambda_i) \tag{1}$$

where λ is the winner model; λ_i is the i^{th} acoustic model and $p(\lambda_i)$ its "a priori" probability; $\mathbf{O} = \{\mathbf{o_1}, \ldots, \mathbf{o_T}\}$ represents the sequence of input feature vectors with $\mathbf{o_t}$ being the observed vector at time instant t, and T the total number of observed vectors; finally, $p(\mathbf{O}|\lambda_i)$ represents the likelihood that the observations $\mathbf{O}$ were generated by the model λ_i.

Taking into account the hidden sequence of states within each model, we rewrite eq. (1) as follows:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} p(\mathbf{o_t}|x_t^i)\right] p(\lambda_i) \tag{2}$$

where $\mathbf{X^i} = \{x_0^i, \ldots, x_T^i\}$ is the state sequence of the i^{th} model that produces the maximum likelihood; $a_{x_t^i x_{t+1}^i}$ refers to the transition probability between the states x_t^i and x_{t+1}^i; and $p(\mathbf{o_t}|x_t^i)$ denotes the likelihood that the observation $\mathbf{o_t}$ was generated in the state x_t^i.

Taking equations (1) and (2) as reference, in the following paragraphs, we

obtain new versions of them for an uncertainty decoding-based recognizer. As mentioned above, uncertainty decoding techniques were proposed for considering some degree of uncertainty in the input feature vectors. The uncertainty of the estimated features, $\hat{\mathbf{O}}$, is modeled through the following conditioned probability distribution:

$$p(\mathbf{O}|\hat{\mathbf{O}}), \tag{3}$$

which models the likelihood that the contaminated observation $\hat{\mathbf{O}}$ came from the clean one ($\mathbf{O}$). Thus, the uncertainty decoding methods rewrite eq. (1) to take into account all the clean observations that could have produced the noisy one (Morris et al. (2001); Yoma and Villar (2002); Deng et al. (2002); Arrowood and Clements (2002); Droppo et al. (2002); Benitez et al. (2004); Stouten et al. (2006)):

$$\begin{aligned}\lambda &= \arg\max_i E\left\{p(\lambda_i)p(\mathbf{O}|\lambda_i)|\mathbf{O}\sim p(\mathbf{O}|\hat{\mathbf{O}})\right\} = \\ &= \arg\max_i p(\lambda_i)\int_{\mathbf{O}} p(\mathbf{O}|\lambda_i)p(\mathbf{O}|\hat{\mathbf{O}})d\mathbf{O}\end{aligned} \tag{4}$$

Assuming now that the observation generated at time t does not depend on other time instants, we write $p(\mathbf{O}|\hat{\mathbf{O}})$ as follows:

$$p(\mathbf{O}|\hat{\mathbf{O}}) = \prod_{t=1}^{T} p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t}) \tag{5}$$

Substituting eq. (5) in eq. (4) and making explicit the state sequence, we obtain:

$$\lambda = \arg\max_i a_{x_0^i x_1^i}\left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i}\int_{\mathbf{o_t}} p(\mathbf{o_t}|x_t^i)p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})d\mathbf{o_t}\right]p(\lambda_i) \tag{6}$$

Now, comparing the decision rule corresponding to a conventional recognizer, eq. (2), and that of an uncertainty decoding-based one, eq. (6), we observe that the term

$$\int_{\mathbf{o_t}} p(\mathbf{o_t}|x_t^i)p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})d\mathbf{o_t} \tag{7}$$

is applied instead of

$$p(\mathbf{o_t}|x_t^i). \tag{8}$$

In Morris et al. (2001); Deng et al. (2002); Krisjansson and Frey (2002) this theory is applied to ASR systems that use log filter bank energies as front-end. Morris et al. (2001) does not use any feature enhancement technique and combines uncertainty and conventional decoding, and the weights of the

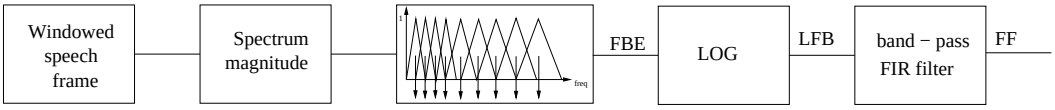

Figure 1. Block diagram of the Frequency Filtered parameterization.

combination are computed as a function of the distortion degree of each observation. Deng et al. (2002) and Krisjansson and Frey (2002) employ feature enhancement techniques. In the first case, the uncertainty is modeled by means of a Gaussian distribution, while in the second case, no one particular distribution is assumed. The ASR systems using log filter bank energies obtain worse results than those using cepstral parameters. Yoma and Villar (2002); Arrowood and Clements (2002); Droppo et al. (2002); Benitez et al. (2004); Stouten et al. (2006) apply uncertainty decoding to cepstrum-based ASR systems for different enhancement techniques. In any case, a clear conclusion is extracted from these works: the parameter enhancement methods are not perfect and, consequently, an uncertainty exists about the estimated parameters that should be considered in the recognizer.

3 Uncertainty decoding with Frequency Filtered parameters

In this section, we describe our proposal, consisting of applying concepts of uncertainty decoding to an ASR system which uses the Frequency Filtered parameterization that has been enhanced by spectral subtraction.

To this purpose, we start reviewing the FF parameterization and studying the effects that the additive noises have on it. Then, we model these effects by means of a Gaussian uncertainty distribution and, finally, we modify the decoding algorithm to incorporate this uncertainty.

3.1 Frequency Filtered parameters, additive noise and Spectral Subtraction

Figure 1 summarizes by means of a block diagram the steps involved in the FF coefficient computation. As can be seen, the substitution of the DCT (Discrete Cosine Transform) by a simple band-pass filter is the main difference with respect to MFCCs. This band-pass filter usually takes the form $z - z^{-1}$ and, therefore, just two log filter bank energies are involved.

In the following, we assume that the speech signal is contaminated with additive and uncorrelated noise. Therefore, the noise and speech components are also additive in the power spectrum domain. Although this property is lost in the magnitude spectrum domain, if either the noise or the speech signal dominates the summation, it can be assumed that this additive property still

holds in the magnitude domain. Since the filter bank energies are just a linear combination of the magnitude spectrum components, this property also holds in the filter bank energy domain. Therefore, we have that

$$\widehat{FBE}_k \approx FBE_k + n_k, \tag{9}$$

where FBE_k and $\widehat{FBE}_k$ are the k^{th} filter bank energies of the clean and noisy speech, respectively, and n_k is the additive noise component of the k^{th} filter bank energy. Let us also assume that each noise component, n_k, is a random variable with mean μ_{n_k} and variance $\sigma^2_{n_k}$.

Considering now the use of spectral subtraction at the front-end, the noise spectrum estimate is removed from the noisy speech spectrum. As a result, the mean of the random variable that models the noise component in each energy band, n_k, is assumed to be zero, i.e., ($\mu_{n_k} = 0$).

Next, the log filter bank energies are computed:

$$\widehat{LFB}_k = \log(\widehat{FBE}_k) \approx \log(FBE_k + n_k). \tag{10}$$

The first order Taylor series expansion of the *log* operator around a certain point a is used to obtain a linear approximation of the log filter bank energies. Therefore, the noisy and clean versions of the log filter bank energies can be written as:

$$\widehat{LFB}_k \approx \log(a) + \frac{FBE_k}{a} - 1 + \frac{n_k}{a} \tag{11}$$

$$LFB_k \approx \log(a) + \frac{FBE_k}{a} - 1. \tag{12}$$

Now, combining equations (11) and (12) we obtain

$$\widehat{LFB}_k \approx LFB_k + \frac{n_k}{a}. \tag{13}$$

To make this approximation accurate, the point a should be close to $FBE_k + n_k$ and FBE_k. The former value is just the SS-based estimate of the later, which is unknown. Therefore, we select $a = FBE_k + n_k$. It is worth noting that (see eq. (13)) the amount of noise at the k^{th} log filter bank energy is inversely proportional to a, i.e., to FBE_k. As a result, the high-energy bands (spectral peaks) are less sensitive to noise than the low-energy bands (spectral valleys). This fact is due to the *log* operator: for high energies, the derivative of the *log* is small and, therefore, less sensitive to variations due to the noise; however, for low energy regions, the derivative of the *log* is higher and more sensitive to the noise.

As we will see in the next section, the log approximation used in this work is mainly used to detect and avoid the contribution of the samples that are

dominated by noise. The approximation in eq. (13) fulfills this objective and avoids the use of more complex approximations such as, for example, VTS (Vector Taylor Series) (Moreno et al. (1996)).

Once we have written $\widehat{LFB}_k$ as a function of LFB_k (eq. (13)), it is easy to write the noisy FF parameters as a function of the clean FF parameters. The FF coefficient for the k^{th} log filter bank energy is then (Vicente-Peña et al. (2006a)):

$$\widehat{FF}_k = \widehat{LFB}_{(k+1)} - \widehat{LFB}_{(k-1)} \approx LFB_{k+1} + \frac{n_{k+1}}{a} - LFB_{k-1} - \frac{n_{k-1}}{b}, \quad (14)$$

with $a = FBE_{k+1} + n_{k+1}$ and $b = FBE_{k-1} + n_{k-1}$. Given that $FF_k = LFB_{(k+1)} - LFB_{(k-1)}$, eq. (14) can be rewritten as:

$$\widehat{FF}_k \approx FF_k + \frac{n_{k+1}}{a} - \frac{n_{k-1}}{b} \quad (15)$$

and finally, denoting

$$N_k = \frac{n_{k+1}}{a} - \frac{n_{k-1}}{b}. \quad (16)$$

we obtain:

$$\widehat{FF}_k \approx FF_k + N_k. \quad (17)$$

Therefore, the random variable N_k determines the uncertainty of the FF parameters. Assuming that the noise components are uncorrelated among each other, the mean and variance of N_k can be computed as follows:

$$\mu_{N_k} = 0 \quad (18)$$

$$\sigma^2_{N_k} = \frac{\sigma^2_{n_{k+1}}}{a^2} + \frac{\sigma^2_{n_{k-1}}}{b^2}. \quad (19)$$

This last assumption does not hold for any noise (in fact, it will depend on the frequency and time structure of the noise signal). Nevertheless, it becomes a good trade-off between analytical simplicity and experimental performance. Once we know how additive noises affect the static FF parameters and assuming that the noise components at different time instants are uncorrelated, it is straightforward to study the effects on the dynamic parameters.

3.2 Modeling the uncertainty of the FF parameters through a Gaussian model

Though several models have been proposed in the literature, there is no good solution for any ASR task. We have chosen to use a Gaussian distribution for modeling the uncertainty of the FF parameters since it is the most common

one, works well for many cases, and leads to simple analytical solutions. Additionally, we conducted some experiments using a uniform distribution but the results turned out to be similar.

Therefore, we assume that the distribution that models the noise component in the FF domain (eq. (17)) is Gaussian:

$$N_{kt} \sim \mathcal{N}(\cdot; 0, \sigma^2_{N_{kt}}) \tag{20}$$

where the mean and variance are given by equations (18) and (19), respectively, and we have added the time index to emphasize the time dependency.

From this hypothesis, in the following paragraphs, we deduce the expression that models the uncertainty of the observations at the front-end, $p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})$ in eq. (6). Following the notation of the previous section, the vector $\mathbf{o_t}$ represents the FF parameters obtained from the clean speech and $\hat{\mathbf{o}}_\mathbf{t}$ represents the estimated parameters after applying spectral subtraction in the magnitude spectrum domain. That is,

$$\mathbf{o_t} = \begin{bmatrix} FF_{1t} \\ \cdots \\ FF_{kt} \\ \cdots \\ FF_{Nt} \end{bmatrix}; \hat{\mathbf{o}}_\mathbf{t} = \begin{bmatrix} \widehat{FF}_{1t} \\ \cdots \\ \widehat{FF}_{kt} \\ \cdots \\ \widehat{FF}_{Nt} \end{bmatrix} \tag{21}$$

where N refers to the dimension of the input feature vector. Formally, we should add the log-energy and the dynamic parameters since the uncertainty of the dynamic parameters has been taken into account in our experiments (the front-end is described in detail in Section 5.1). Nevertheless, for the sake of clarity, we have not explicitly included them in the formulas.

Assuming that the FF coefficients are uncorrelated with each other, the term $p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})$ is rewritten as:

$$p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t}) = \prod_{k=1}^{N} p(FF_{kt}|\widehat{FF}_{kt}) \tag{22}$$

where $p(FF_{kt}|\widehat{FF}_{kt})$, the probability distribution that models the uncertainty of each FF component, can be obtained from equations (17) and (20) as follows:

$$p(FF_{kt}|\widehat{FF}_{kt}) = \mathcal{N}(FF_{kt}; \widehat{FF}_{kt}, \sigma^2_{N_{kt}}) \tag{23}$$

Once $p(\mathbf{o_t}|\hat{\mathbf{o}}_\mathbf{t})$ has been determined, the decision rule of the recognizer based

on uncertainty decoding can easily be obtained. First, we assume that the model distribution used in the HMMs is a Gaussian mixture:

$$p(\mathbf{o_t}|x_t^i) = \sum_{m=1}^{M} c_{x_t^i m} \mathcal{N}(\mathbf{o_t}; \boldsymbol{\mu}_{\boldsymbol{x_t^i m}}, \boldsymbol{\Sigma}_{\boldsymbol{x_t^i m}}) \tag{24}$$

where M refers to the number of Gaussians per state; $\boldsymbol{\mu}_{\boldsymbol{x_t^i m}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x_t^i m}}$ refers to the mean vector and covariance matrix in mixture m and state x_t^i, respectively.

Considering now diagonal matrices, the above equation becomes:

$$p(\mathbf{o_t}|x_t^i) = \sum_{m=1}^{M} c_{x_t^i m} \prod_{k=1}^{N} \mathcal{N}(FF_{kt}; \mu_{x_t^i mk}, \sigma^2_{x_t^i mk}) \tag{25}$$

where $\mu_{x_t^i mk}$ is the k^{th} component of the vector $\boldsymbol{\mu}_{\boldsymbol{x_t^i m}}$ and $\sigma^2_{x_t^i mk}$ the k^{th} component of the diagonal in the matrix $\boldsymbol{\Sigma}_{\boldsymbol{x_t^i m}}$.

Using equations (23) and (25), the decision rule given by eq. (6) can be rewritten as follows:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \sum_{m=1}^{M} c_{x_t^i m} \left\{ \prod_{k=1}^{N} \int_{FF_{kt}} \mathcal{N}(FF_{kt}; \mu_{x_t^i mk}, \sigma^2_{x_t^i mk}) \mathcal{N}(FF_{kt}; \widehat{FF}_{kt}, \sigma^2_{N_{kt}}) dFF_{kt} \right\} \right] \tag{26}$$

Considering now that the second Gaussian can be rewritten as follows:

$$\mathcal{N}(FF_{kt}; \widehat{FF}_{kt}, \sigma^2_{N_{kt}}) = \mathcal{N}(FF_{kt} - \widehat{FF}_{kt}; 0, \sigma^2_{N_{kt}}) = \mathcal{N}(\widehat{FF}_{kt} - FF_{kt}; 0, \sigma^2_{N_{kt}}), \tag{27}$$

the integral in eq. (26) can be expressed as the next convolution operation that is evaluated at the point $\widehat{FF}_{kt}$:

$$\begin{aligned} &\int_{FF_{kt}} \mathcal{N}(FF_{kt}; \mu_{x_t^i mk}, \sigma^2_{x_t^i mk}) \mathcal{N}(FF_{kt}; \widehat{FF}_{kt}, \sigma^2_{N_{kt}}) dFF_{kt} = \\ &= \left[\mathcal{N}(FF_{kt}; \mu_{x_t^i mk}, \sigma^2_{x_t^i mk}) * \mathcal{N}(FF_{kt}; 0, \sigma^2_{N_{kt}}) \right]_{FF_{kt} = \widehat{FF}_{kt}} \end{aligned} \tag{28}$$

where $*$ denotes the convolution operator. The result of the convolution is a new Gaussian whose mean is the sum of the individual means and whose variance is the sum of the individual variances. Thus, eq. (26) can be rewritten as:

$$\lambda = \arg\max_i a_{x_0^i x_1^i} \left[\prod_{t=1}^{T} a_{x_t^i x_{t+1}^i} \sum_{m=1}^{M} c_{x_t^i m} \prod_{k=1}^{N} \mathcal{N}(\widehat{FF}_{kt}; \mu_{x_t^i mk}, \sigma^2_{x_t^i mk} + \sigma^2_{N_{kt}}) \right] \tag{29}$$

As we observe in the above equation, the new decision rule with uncertainty decoding consists just of adding a time-variant noise-dependent term, $\sigma^2_{N_{kt}}$, to the original variance.

Thereby, the modified decision rule can be interpreted as a simple variance adaptation method. Indeed, if we compute the mean and variance of the estimated parameters:

$$\begin{aligned}\mu_{\widehat{FF}_{kt}} &= E\{\widehat{FF}_{kt}\} = E\{FF_{kt}\} = \mu_{x_t^i mk},\\ \sigma^2_{\widehat{FF}_{kt}} &= E\{(\widehat{FF}_{kt} - \mu_{x_t^i mk})^2)\} = \sigma^2_{x_t^i mk} + \sigma^2_{N_{kt}},\end{aligned} \tag{30}$$

we observe that the mean of the estimated parameters is equal to the mean of the original parameters, but the variance has changed and should be adapted.

Finally, we have considered in our experiments an extension of the Gaussian distribution of eq. (20) by using a scaled standard deviation through a parameter δ:

$$N_{kt} \sim \mathcal{N}(\cdot; 0, (\delta\sigma_{N_{kt}})^2) \tag{31}$$

The use of this parameter allows us to compensate to some extent for inaccuracies in the Gaussian model assumed by eq. (20).

3.2.1 Interpretation of the Gaussian model of the uncertainty as a spectral weighting

Looking at the uncertainty decoding using a Gaussian model as a variance adaptation method allows us to make a new interpretation based on spectral weighting (Vicente-Peña et al. (2006a)). To this end, we need to recall the conventional decision rule, eq. (2), in terms of log-likelihoods:

$$\lambda = \arg\max_i \left(\log(a_{x_0^i x_1^i}) + \sum_{t=1}^{T} \left[\log(a_{x_t^i x_{t+1}^i}) + \log(p(\mathbf{o_t}|x_t^i)) \right] + \log(p(\lambda_i)) \right). \tag{32}$$

Now, we focus on the term that depends on the emission probability at each state, $p(\mathbf{o_t}|x_t^i)$, suppressing the time index dependence for the sake of clarity:

$$\log(p(\mathbf{o}|x_t^i = j)) = -\frac{1}{2}\left\{ \sum_{k=1}^{N} \log(2\pi\sigma_{jk}^2) + \sum_{k=1}^{N} \frac{(FF_k - \mu_{jk})^2}{\sigma_{jk}^2} \right\} \tag{33}$$

where we have considered the state distribution as a single Gaussian whose mean and variance for the k^{th} component are μ_{jk} and σ_{jk}^2, respectively, and

we have denoted the state x_t^i by j for simplicity. The general case of mixture of Gaussians can be developed in a similar way, but we have preferred to keep the equations simpler.

If we adapt the model variance to take into account the uncertainty that is present in the estimated parameters, eq. (33) becomes:

$$\log(p(\hat{\mathbf{o}}|x_t^i = j)) = -\frac{1}{2}\left\{\sum_{k=1}^{N}\log\left(2\pi(\sigma_{jk}^2 + \sigma_{N_k}^2)\right) + \sum_{k=1}^{N}\frac{(\widehat{FF}_k - \mu_{jk})^2}{\sigma_{jk}^2 + \sigma_{N_k}^2}\right\} \quad (34)$$

Using the notation,

$$w_{jk} = \frac{\sigma_{jk}^2}{\sigma_{jk}^2 + \sigma_{N_k}^2}, \quad (35)$$

and rewriting eq. (34) in a more convenient way, we obtain:

$$\log(p(\hat{\mathbf{o}}|x_t^i = j)) = -\frac{1}{2}\left\{\sum_{k=1}^{N}\log(2\pi\sigma_{jk}^2) + \sum_{k=1}^{N}w_{jk}\frac{(\widehat{FF}_k - \mu_{jk})^2}{\sigma_{jk}^2} - \sum_{k=1}^{N}\log w_{jk}\right\}. \quad (36)$$

If we compare this equation with the one that corresponds with the criterion followed for the clean parameters (eq. (33)) two differences become evident:

- The term
$$\sum_{k=1}^{N}\frac{(FF_k - \mu_{FF_k})^2}{\sigma_{jk}^2} \quad (37)$$
for clean features turns into
$$\sum_{k=1}^{N}w_k\frac{(FF_{kn} - \mu_{jk})^2}{\sigma_{jk}^2} \quad (38)$$
for noise features. This term is a normalized Euclidean distance that indicates how near or far is the observation from the model represented by $(\mu_{jk}, \sigma_{jk}^2)$.

  We can see the weights w_k given by eq. (35) as a measure of the noise level in our input features. Therefore, when the noisy term in our variance, $\sigma_{n_{k+1}}^2/a^2 + \sigma_{n_{k-1}}^2/b^2$, is relatively small, we find weights close to one. In contrast, the weights are close to zero when the noisy term is relatively large. Normally, weights close to one come from high-energy regions in the log filter bank energy domain and, therefore, eq. (38) is dominated by the spectral peaks instead of by the valleys. It is also important to note that the weights depend on the variance of the model in such a way that models with larger variances are less sensitive to noise distortions.
- The second difference consists of the addition of the term
$$-\sum_{k=1}^{N}\log w_k. \quad (39)$$

The problem when we weight the Euclidean distance as we do in eq. (38) is that it is close to zero if most of the weights are low. A distance close to zero indicates a perfect matching between the current model and the observation. The term defined by eq. (39) adds a penalty for low weights. It is worth noting that this term vanishes when all weights are equal to one, that is, when there is no noise.

4 SSBD-HMM

In order to provide a suitable background for describing the combination of SSBD-HMM and uncertainty decoding, a brief review of the underlying ideas of the SSBD-HMM method are given in this section. For more details, the reader is referred to Vicente-Peña et al. (2010).

An HMM-based speech recognizer computes the log-likelihood of the sequence of observed features for every candidate acoustic model and selects the candidate that provides the maximum. This computation entails the calculation of emission log-probabilities, which depends on the normalized Euclidean distance between the current observation and the corresponding model mean. Some of these normalized Euclidean distances can dominate the log-probabilities computation when one of their components is strongly corrupted (i.e., it is an outlier) and, consequently, is far away from its corresponding mean. Thus, the corrupted components contribute strongly towards discarding the corresponding model. For this reason, it is convenient to bound the distance and, as a result, to bound the influence of the outliers on the final decision.

The BD-HMM method suggests substituting the normalized Euclidean distance by a bounded distance as follows:

$$BD(a-b) = \begin{cases} (a-b)^2, \text{ if } |a-b| < B \\ B^2, \qquad \text{otherwise} \end{cases} \tag{40}$$

where, in our case, a is a component of the observation vector, b is the corresponding component of the mean vector, and B is a bound directly related to the corresponding component of the variance vector (assuming diagonal covariance matrices).

The BD-HMM method is effective to cope with outliers that appear when ASR systems deal with noisy speech. However, the BD-HMM method works only on the outliers, leaving the remaining effects of noise. On the other hand, SS tries to estimate the clean parameters from the noise-corrupted ones and, unlike BD-HMM, works on all the observed features. SS was not originally

designed as a preprocessing stage for speech recognition but as a speech enhancement method. In this context, it is well known that the noise reduction is attained at the expense of introducing distortions. As shown in Vicente-Peña et al. (2010), these distortions produce an increment of the number of outliers. Taking this fact into consideration, the joint application of SS and BD-HMM, called SSBD-HMM, takes advantage of the speech enhancement attained by SS while avoiding, by means of the BD-HMM, the negative effect derived from the SS-generated outliers in the estimated spectral densities. Furthermore, the BD-HMM method will be complemented by SS, which compensates for all the parameters (not just the outliers).

5 Experiments and Results

5.1 System set-up

The experiments were performed using two well-known databases, namely: RM1 NIST (1992) and Aurora-4 Hirsch (2002). For each database we designed a speech recognition system based on the HTK toolkit (Young et al. (2002)).

The same parameterization was used for both databases. Specifically, the input feature vector consisted of 12 FF parameters as well as the log-energy coefficient that were obtained every 10 ms using a 25 ms Hamming window. The time mean of the FF parameters along each utterance was removed, the log-energy was normalized, and the resulting feature vector was extended with the first and second time derivatives, resulting in a 39-dimensional vector.

The magnitude of the spectrum was estimated by applying spectral subtraction to the noise-contaminated utterances. The version of spectral subtraction used in the experiments follows the next equation:

$$|\widehat{X}(\omega,l)| = \max\left\{\left(|\widehat{S}(\omega,l)| - \gamma|\widehat{N}(\omega,l)|\right), \beta|\widehat{N}(\omega,l)|\right\} \tag{41}$$

where $|\widehat{X}|$, $|\widehat{S}|$ and $|\widehat{N}|$ are, respectively, the estimates of the clean, noisy speech and noise magnitude spectra; ω is the frequency index and l is the time (frame) index; finally, γ and β are design constants respectively known as "over-estimation factor" and "spectrum flooring." For our experiments, we used $\gamma = 0.8$ and $\beta = 0.2$. $|\widehat{S}|$ was estimated as an average over the magnitude spectra at the previous, current, and next time frames (Boll (1979)), while $|\widehat{N}|$ was estimated using the minimum statistics method proposed in Martin (2001). In order to estimate the noise power spectrum density, Martin (2001) assumes that each noise power spectrum component follows an Exponential

distribution. However, since we were interested in estimating the noise magnitude instead of the noise power spectrum, we used a Rayleigh model, resulting from applying a root square operator on an Exponential variable (Papoulis and Pillai (2002)). In particular, the mean and variance of this Rayleigh random variable are related to the mean of the Exponential random variable by means of the next equations:

$$\mu_{RAY} = \frac{\sqrt{\pi}}{2}\sqrt{\mu_{EXP}} \tag{42}$$

$$\sigma^2_{RAY} = \left(1 - \frac{\pi}{4}\right)\mu_{EXP} \tag{43}$$

where μ_{RAY} and σ^2_{RAY} are, respectively, the mean and variance of the Rayleigh random variable; and μ_{EXP} refers to the mean of the Exponential random variable that represents the mean of the noise power spectrum density. From eq. (42) and (43), it is straightforward to infer the mean and variance of the noise components at the filter-bank energy domain.

Finally, we combined our proposal with the SSBD-HMM method (Vicente-Peña et al. (2010)). As SSBD-HMM employs SS to get clean parameter estimates, we incorporate the information about the uncertainty that remains in the estimates for making the system more effective.

In the next subsections, we give more details of the designed systems that are specific to each database.

5.1.1 Resource Management RM1 (NIST (1992))

The well-known Resource Management RM1 database has a vocabulary of 991 words. The training corpus consists of 3990 sentences, and the test set, which corresponds to a compilation of the first four official test sets, contains 1200 sentences. We used a down-sampled version (at 8 kHz) of the database (originally recorded at 16 kHz in clean conditions).

The orthographic transcription of the data is based on the SRI Resource Management dictionary (provided in the same distribution by NIST). Context-dependent acoustic models were used (cross-word triphones). A three-state, three-mixture per state HMM was used to model each triphone. Two silence models, long and short, were used. The long silence model consisted of three states while the short silence model consisted of a unique state tied to the middle state of the long silence model. Finally, standard word-pair grammar was used as the language model.

Artificially contaminated versions of this database were created by adding four kinds of noises at four different SNRs. Specifically, 8 kHz down-sampled

versions of the pink, car, babble, and factory noises from the NOISEX-92 (Varga et al. (1992)) database were used. The considered SNR values went from 0 dB to 15 dB in 5 dB steps. These contaminated versions were only used for testing purposes (never for training).

5.1.2 Aurora-4 (Hirsch (2002))

This database is based on the WSJ0 database (Paul and Baker (1992)). As a training set, we chose a predefined set, which comprises the clean speech sentences acquired with the close-talking microphone. This set agrees with the standard SI-84 training set defined for the WSJ0, which contains 7138 utterances. The test set was based on the Nov'92 development set, defined for the WSJ0 database that includes artificially noise-contaminated sentences. The noisy versions were created by adding one type of noise at a randomly chosen SNR between 5 and 15 dB in steps of 1 dB. Six different kinds of noises were used: car, babble, restaurant, street, airport, and train station. We used all the noises for our experiments but we limited the experiments to 8 kHz down-sampled versions of the close-talking microphone set.

The CMU dictionary (v 0.6) (CMU (1998)) was used, where we removed the vowel stress obtaining 39 phonemes for our transcriptions. As we did for the RM1 database, cross-word triphone acoustic models and two different silence models were used. In this case, three-state models with 8 Gaussians per state were used, except for the silence models, in which the number of Gaussians was increased to 16. Finally, the 5K bigram language model distributed with the WSJ0 corpus was used.

5.2 Results

Before presenting the results for the noisy experiments, recognition results in terms of Word Error Rate (WER) for clean speech are shown in Table 1. These results are similar to those achieved by other systems that use the same database but different parameterization (e. g. Vicente-Peña et al. (2006b), Woodland et al. (1994) or Parihar and Picole (2001)). In the remainder of this section, we show the performance of our proposals for each database under noisy conditions.

5.2.1 Results for the RM1 task

Figure 2 shows the WER for the RM1 database with a Gaussian distribution for modeling the uncertainty of the parameters (labeled as *UD-G*). In this first experiment, we used $\delta = 1$ in eq. (31). In the figure, we also show the

Table 1
WER (%) for clean speech attained by each of the designed systems for the RM1 and Aurora-4 databases.

	Baseline WER (%)
RM1	6.57 %
Aurora-4	8.8 %

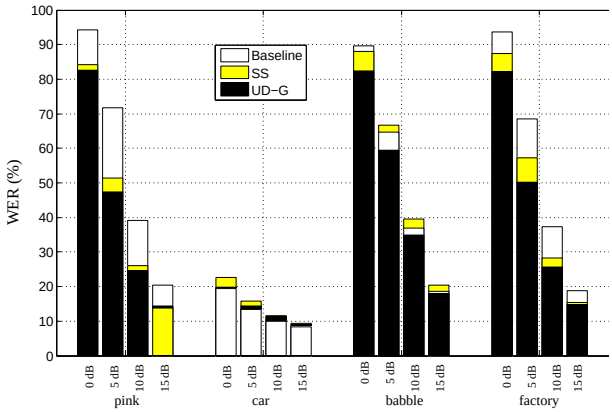

Figure 2. WER for the RM1 task: reference system (*Baseline*), spectral subtraction (*SS*) and a Gaussian model of the uncertainty (*UD-G*).

performance of the baseline system with and without applying SS (labeled, as *SS* and *Baseline*, respectively). As observed in the figure, the effectiveness of uncertainty decoding is clear except for the car noise, for which the SS is not working properly. SS does not work for babble noise either (except at 0 dB), but in this case, uncertainty decoding is able to compensate for the losses owing to SS. In the remaining cases, the situation is as expected: SS improves the results achieved by the baseline system and uncertainty decoding improves even more the results achieved by SS. In particular, the improvements over SS are statistically significant [1] for pink noise at 0 and 5 dB; for factory noise at 0, 5 and 15 dB and, finally, for babble noise, the improvements are significant for all the studied SNRs. In summary, the results are statistically significant for low and medium SNRs.

Figure 3 compares the results achieved by uncertainty decoding with those achieved by SSBD-HMM. The results clearly prove that SSBD-HMM is more effective dealing with additive noises than uncertainty decoding. However, uncertainty decoding improves the way of dealing with those contaminated features that are not outliers and, therefore, we suggest combining SSBD-HMM with uncertainty decoding. In the same Figure 3, we show the results for this combination (labeled as *UDBD-G*). The combination generally achieves the

[1] We have stated the statistical significance of the results calculating the confidence intervals, for a confidence of 95 % (see Weiss and Hasset (1993), for details).

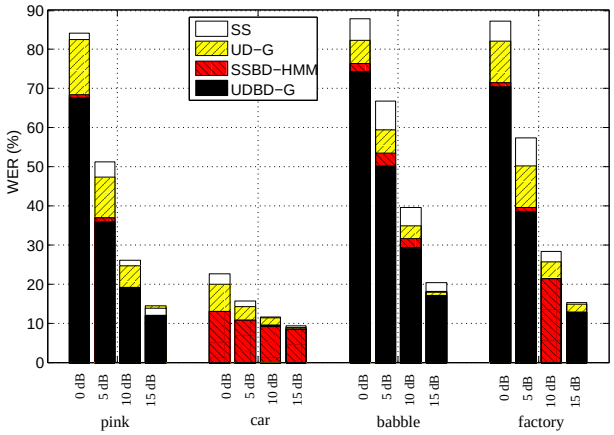

Figure 3. WER for the RM1 database: spectral subtraction (*SS*), uncertainty decoding using a Gaussian distribution (*UD-G*), combination of SS and BD-HMM (*SSBD-HMM*) and combination of uncertainty decoding using a Gaussian distribution and SSBD-HMM (*UDBD-G*).

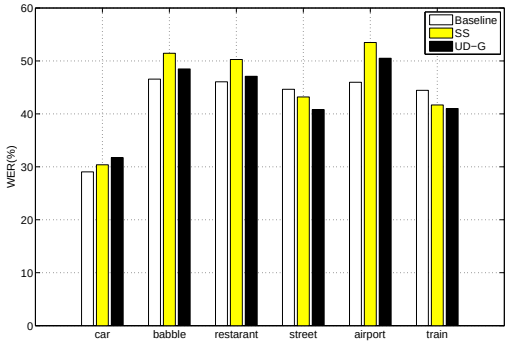

Figure 4. WER for the Aurora-4 database: uncertainty decoding with Gaussian distribution (*UD: Gauss.*), spectral subtraction (*SS*) and the reference system (*Baseline*).

best results, with the improvements respect to SS being statistically significant for all the cases with the exception of car noise at 15 dB. Furthermore, the improvement over SSBD-HMM due to uncertainty decoding is statistically significant for babble noise at 0,5 and 10 dB.

For these experiments, no significant differences were observed by using several values of δ in eq. (31), that was varied between 0.75 and 2.0. We used $\delta = 1.0$.

5.2.2 Results for the Aurora-4 task

Figure 4 shows the results achieved by uncertainty decoding for the Aurora-4 database. With the exception of car noise, uncertainty decoding outperforms SS with the improvements being statistically significant for babble, restaurant, and airport noises. However, the baseline system generally achieves the best

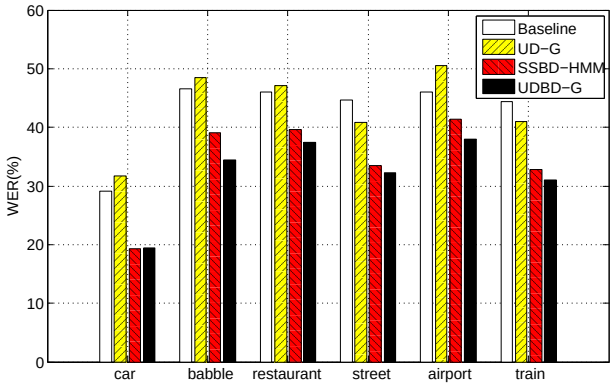

Figure 5. WER for the Aurora-4 database: comparison between uncertainty decoding and SSBD-HMM.

results because SS does not work properly with four out of six types of noise.

For the same reason explained previously, we suggest combining uncertainty decoding with SSBD-HMM. As can be observed in Figure 5, the suggested combination achieves the best results in all the cases except for car noise. Furthermore, all of the improvements with respect to the Baseline or SS are statistically significant.

The fact that the proposed combination achieves the best results (with the mentioned exception) also proves that uncertainty decoding clearly contributes to improving the already good results achieved by SSBD-HMM. We should point out that car noise is mainly stationary and the uncorrelation property assumed in sec. 3.1 for estimating the uncertainty in the dynamic parameters is likely not as precise as required. For the remaining noises, uncertainty decoding combined with SSBD-HMM obtains a relative improvement with respect to SSBD-HMM equal to 12 % for babble noise, 8 % for airport noise, around 5 % for train and restaurant noises and 3 % for street noise. These improvements are statistically significant for babble and airport noises.

Again, several values for the δ parameter were tested, namely: from $\delta = 0.75$ to $\delta = 2.0$ in steps of 0.25. The best results were found for $\delta = 1.75$. It is worth noting that this value is larger than the one used for the RM1 task. This difference is likely due to the type of noises considered in the Aurora-4 database, which makes the estimation of the clean parameters more difficult. As a result, the application of spectral subtraction leaves a higher level of uncertainty. In any case, recognition rates are not very sensitive to small variations of the δ parameter.

6 Conclusions

The method proposed in this paper starts from a system that mitigates the effects of additive noises by means of spectral subtraction. We applied uncertainty decoding to include into the recognition process the information about the uncertainty that still remains in the observations after SS. Specifically, our work focused on FF parameterization, which achieves as good results as the well-known MFCC parameterization but has the advantage of remaining in the log-spectrum domain. Taking advantage of this fact, we obtained simple equations for describing the effects of additive noises in the FF domain. These simple relations allowed us to model the uncertainty present at the front-end. We used a Gaussian distribution for modeling this uncertainty, inferring the new decision rule that governs the recognition process. Additionally, this new decision rule allowed us to develop a novel interpretation of the uncertainty decoding method as a spectral weighting technique. Our first results showed the effectiveness of the uncertainty decoding.

In addition, we suggested combining the proposed uncertainty decoding method with an effective technique for dealing with additive noise, known as SSBD-HMM. The experimental results on the well-known RM1 and Aurora-4 ASR tasks clearly show how uncertainty decoding significantly improves the SSBD-HMM performance. This improvement is explained by the fact that uncertainty decoding methods are able to compensate for features that, without being outliers for the models of the recognizer, are also affected by the presence of noise.

Acknowledgements

The authors would like to thank Prof. Bastiaan Kleijn for his support along this work.

References

Arrowood, J. A., Clements, M. A., Sep. 2002. Using observation uncertainty in HMM decoding. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP). pp. 1561–1564.

Benitez, C., Segura, J. C., de la Torre, A., Ramirez, J., Rubio, A. J., Oct. 2004. Including uncertainty of speech observations in robust speech recognition. In: Proc. Int. Conf. on Spoken Language Processing (INTERSPEECH - ICSLP). pp. 137–140.

Boll, S. F., Apr. 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics, Speech, Signal Processing 27 (2), 113–120.

CMU, 1998. The CMU (v 0.6) pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

de Veth, J., Cranen, B., Boves, L., Jun. 2001a. Acoustic backing-off as an implementation of missing feature theory. Speech Communication 34 (3), 247–265.

de Veth, J., de Wet, F., Cranen, B., Boves, L., Apr. 2001b. Acoustic features and a distance measure that reduce the impact of training-test mismatch in ASR. Speech Communication 34 (1), 57–74.

Deng, L., Droppo, J., Acero, A., Sep. 2002. Exploiting variances in robust feature extraction based on a parametric model of speech distortion. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP). pp. 2449–2452.

Droppo, J., Acero, A., Deng, L., May 2002. ncertainty decoding with SPLICE for noise robust speech recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP). Vol. 1. pp. 57–60.

Gong, Y., Apr. 1995. Speech recognition in noisy environments: a survey. Speech Communication 16 (3), 261–291.

Hirsch, G., Nov. 2002. Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task, Version 2.0, AU/417/02. Tech. rep., ETSI STQ-Aurora DSR Working Group.

Krisjansson, T. T., Frey, B. J., May 2002. Accounting for uncertainty in observations: a new paradigm for robust automatic speech recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP). Vol. 1. pp. 61–64.

Martin, R., Jul. 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Processing 9 (5), 504–512.

Moreno, P., Raj, B., Stern, R., May 1996. A vector Taylor series approach for environment-independent speech recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP). Vol. 2. pp. 733–736.

Morris, A., Barker, J., , Bourlard, H., apr 2001. From missing data to maybe useful data: soft data modelling for noise robust ASR. In: WISP workshop on innovative methods in speech recognition.

Nadeu, C., Hernando, J., Gorricho, M., Sep. 1995. On the decorrelation of filter-bank energies in speech recognition. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech). pp. 1381–1384.

Nadeu, C., Macho, D., Hernando, J., Apr. 2001. Time and frequency filtering of filter-bank energies for robust HMM speech recognition. Speech Communication 34 (1-2), 93–114.

NIST, 1992. NIST, The Resource Management Corpus(RM1). Distributed by NIST.

Paliwal, K. K., Sep. 1999. Decorrelated and liftered filter-bank energies for

robust speech recognition. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech). pp. 85–88.

Papoulis, A., Pillai, S. U., 2002. Probability, Random Variable and Stochastic Processes, 4th Edition. McGraw-Hill.

Parihar, N., Picole, J., Jul. 2001. Aurora Working Group: DSR Front End LVCSR Evaluation - baseline recognition system description. Tech. rep., ETSI STQ-Aurora DSR Working Group.

Paul, D. B., Baker, J. M., 1992. The design for the wall street journal-based CSR corpus. In: Human Language Technology Conference. pp. 357–362.

Rabiner, L. R., Feb. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. IEEE 77 (2), 257–286.

Stouten, V., hamme, H. V., Wambacq, P., Nov. 2006. Model-based feature enhancement with uncertainty decoding for noise robust ASR. Speech Communication 48 (11), 1502–1514.

Varga, A. P., Steenneken, J. M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on Automatic Speech Recognition. In: Tech. Rep. DRA Speech Res. Unit. Malvern, Worcestershire, U. K.

Vicente-Peña, J., Díaz-de-María, F., Kleijn, W. B., Sep. 2006a. Individual online variance adaptation of frequency filtered parameters for robust ASR. In: Proc. Int. Conf. on Spoken Language Processing (INTERSPEECH - ICSLP). pp. 1491–1494.

Vicente-Peña, J., Díaz-de-María, F., Kleijn, W. B., Feb. 2010. The synergy between bounded-distance HMM and spectral subtraction for robust speech recognition. Speech Communication 52 (2), 123–133.

Vicente-Peña, J., Gallardo-Antolín, A., Peláez-Moreno, C., de María, F. D., Jul. 2006b. Band-pass filtering of the time sequences of spectral parameters for robust wireless speech recognition. Speech Communication 48 (10), 1379–1398.

Weiss, N. A., Hasset, M. J., 1993. Introductory statistics, 3rd Edition. Addison-Wesley, pp. 407–408.

Woodland, P. C., Odell, J. J., Valtchev, V., Young, S. J., Apr. 1994. Large vocabulary continuous speech recognition using htk. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP). Vol. 2. pp. 125–128.

Yoma, N., McInnes, F., Jack, M., Nov. 1998. Improving performance of spectral subtraction in speech recognition using a model for additive noise. IEEE Trans. Speech Audio Processing 6 (6), 579–582.

Yoma, N. B., Villar, M., Mar. 2002. Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm. IEEE Trans. Speech Audio Processing 10 (3), 158–166.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. The HTK Book (for HTK Version 3.2.1). Cambridge Univ. Press, Cambridge, U.K.