The Synergy between Bounded-Distance HMM and Spectral Subtraction for Robust Speech Recognition

Jesús Vicente-Peña[1], *Fernando Díaz-de-María*[1], *W. Bastiaan Kleijn*[2]

[1] *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
[2] *Sound and Image Processing Lab.*
KTH (Royal Institute of Technology), Stockholm, Sweden

Abstract

Additive noise generates important losses in automatic speech recognition systems. In this paper, we show that one of the causes contributing to these losses is the fact that conventional recognisers take into consideration feature values that are outliers. The method that we call bounded-distance HMM is a suitable method to avoid that outliers contribute to the recogniser decision. However, this method just deals with outliers, leaving the remaining features unaltered. In contrast, spectral subtraction is able to correct all the features at the expense of introducing some artifacts that, as shown in the paper, cause a larger number of outliers. As a result, we find that bounded-distance HMM and spectral subtraction complement each other well. A comprehensive experimental evaluation was conducted, considering several well-known ASR tasks (of different complexities) and numerous noise types and SNRs. The achieved results show that the suggested combination generally outperforms both the bounded-distance HMM and spectral subtraction individually. Furthermore, the obtained improvements, especially for low and medium SNRs, are larger than the sum of the improvements individually obtained by bounded-distance HMM and spectral subtraction.

Key words: Robust speech recognition, spectral subtraction, acoustic backing-off, bounded-distance HMM, missing features, outliers

Email address: jvicente,fdiaz@tsc.uc3m.es; bastiaan.kleijn@ee.kth.se.

1 Introduction

State-of-the-art Automatic Speech Recognition (ASR) systems can achieve high recognition rates in distortion-free environments. However, the differences between the acoustical environment in the real-world application and that used for gathering the training data cause a significant degradation in recognition performance. A significant research effort has been devoted to tackle this *mismatch problem*, particularly for the cases of convolutive distortion and additive noise.

The work presented in this paper focuses on dealing with additive noise. Numerous papers can be found on this subject. The interested reader is referred to the classical paper due to Gong (1995) for an excellent review. The approaches to mitigating the effects of additive noise can be divided into three classes:

- *robust parameterisation*: the selection of a set of robust features that is relatively invariant to additive distortions. Within this category we find techniques such as RASTA-PLP (Hermansky and Morgan (1994)), CMN (Cepstral Mean Normalisation) (Furui (1981)), SCMN (Segmental Cepstral Mean Normalisation) (Viikki and Laurila (1998)), VTLN (Vocal Tract Length Normalisation) (Hain et al. (1999)) or histogram equalization (de la Torre et al. (2005)).
- *feature enhancement*: in this case, the recogniser works with estimates of the clean features that are obtained from the noisy ones. In general, these methods were originally developed for improving the speech quality, but they do not necessarily improve the recognition performance (see Gong (1995)). This second category includes popular methods such as spectral subtraction (Boll (1979)), VTS (Vector Tailor Series) (Moreno et al. (1996)) or SPLICE (Deng et al. (2000)).
- *model compensation*: the third approach entails modifying the acoustic models embedded in the recogniser in order to adapt them to better model the noisy speech. PMC (Gales and Young (1996)) and MLLR (Gales and Woodland (1996)) are good examples of methods within this category.

Recently, so-called missing-feature methods (Cooke et al. (2001); Raj et al. (2004)) have been proposed as a novel technique for robust ASR. They are difficult to classify within one of the forementioned classes. The underlying idea of missing-feature methods is simple: the recogniser uses only the most reliable features. Then, the new problem consists in detecting the unreliable portions of the time-frequency spectrogram and to remove them from the recognition process. Missing-feature approaches have shown a high efficiency when there is perfect knowledge about the reliability of the features. However, the errors due to incorrect feature selection may cause significant loss of performance.

In this paper we suggest a method that we call bounded-distance HMM, which resembles a method already proposed in speaker recognition (Matsui and Furui (1992)). The aim of bounded-distance HMM is to mitigate the influence of the features that are outliers for each acoustic model. Given the apparent relationship of this method with missing-feature approaches, an interpretation of bounded-distance HMM as a missing-feature method is discussed in the paper.

Although the implementation is different, the technique known as acoustic backing-off (de Veth et al. (2001a)) is similar to bounded-distance HMM. Acoustic backing-off adds a uniform distribution to the actual distribution of the acoustic models to model the features not well represented during the training phase. As a result, these unseen features do not play a relevant role in the recogniser decision. As reported in (de Veth et al. (2001b)) the main drawback of this technique is its limited performance for wide-band noises.

As shown in this paper, the combination of bounded-distance HMM with spectral subtraction is able to overcome some of the drawbacks associated with the acoustic backing-off method. Thus, spectral subtraction is an excellent companion method. Essentially, the distortions resulting from spectral subtraction are properly countered by bounded-distance HMM, which, furthermore, takes advantage of the feature enhancement carried out by spectral subtraction.

In this paper, the effectiveness of the combination of bounded-distance HMM and spectral subtraction is theoretically motivated and experimentally proved. The methods are experimentally assessed for several tasks (of different complexities) and for several noise types and signal-to-noise ratios, obtaining very encouraging results.

The paper is organised as follows. Section 2 introduces the bounded-distance HMM method. In Section 3 we propose the combination of bounded distance HMM and spectral subtraction for getting robust systems including some implementation details. Next, in Section 4, we describe the experimental setup and we show and discuss the results achieved by the proposed method. Finally, the main conclusions are outlined in Section 5.

2 Bounded-distance HMM

2.1 Motivation and Antecedents

A HMM-based speech recogniser computes the log-likelihood of the sequence of observed features for every candidate acoustic model and selects the candidate

that provides the maximum (see either Rabiner (1989) or Young et al. (2002) for more details):

$$\lambda = \arg\max_i \left(\log(a^i_{x_o x_1}) + \sum_{t=1}^{L} \left[\log(b^i_{x_t}(\mathbf{o}_t)) + \log(a^i_{x_t x_{t+1}}) \right] + \log(Pr(\lambda_i)) \right) \tag{1}$$

where:

- λ_i is the acoustic model i and λ is the winner acoustic model.
- $a^i_{x_t x_{t+1}}$ is the transition probability between the states x_t and x_{t+1} for the model λ_i.
- $\mathbf{o}_t$ is the observed feature vector at the time instant t.
- $b^i_{x_t}(\mathbf{o}_t)$ is the emission probability for the state x_t of the model λ_i
- L is the number of input feature vectors in the utterance that is being recognised.
- $Pr(\lambda_i)$ is the "a priori" probability of the model λ_i.

In order to motivate the Bounded-Distance HMM (BD-HMM) method we now focus on the emission log-probability, $\log(b^i_{x_t}(\mathbf{o}_t))$. Assuming that this log-probability is modelled by a single Gaussian, this term can be expanded as follows:

$$\log(b^i_{x_t}(\mathbf{o}_t)) = -\frac{1}{2}\log\left((2\pi)^N |\mathbf{\Sigma}^i_{x_t}|\right) - \frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}^i_{x_t})^T (\mathbf{\Sigma}^i_{x_t})^{-1} (\mathbf{o}_t - \boldsymbol{\mu}^i_{x_t}) \tag{2}$$

where N is the feature vector dimension and $\boldsymbol{\mu}^i_{x_t}$ and $\mathbf{\Sigma}^i_{x_t}$ are the mean vector and the covariance matrix, respectively, for the model i and state x_t.

Dropping in equation (2), for simplicity, the time, state and model indexes and considering diagonal covariance matrices, we obtain:

$$\log(b(\mathbf{o})) = -\frac{1}{2}\left\{ \sum_{k=1}^{N} \log(2\pi\sigma_k^2) + \sum_{k=1}^{N} \frac{(o_k - \mu_k)^2}{\sigma_k^2} \right\} \tag{3}$$

where σ_k^2 refers to the kth component of the covariance diagonal matrix and μ_k refers to the kth component of the mean vector.

It is clear that the log-probability exhibits a strong dependence on the normalised euclidean distance between the current observation and the model mean. In particular, the term

$$-\frac{1}{2}\sum_{k=1}^{N} \frac{(o_k - \mu_k)^2}{\sigma_k^2} \tag{4}$$

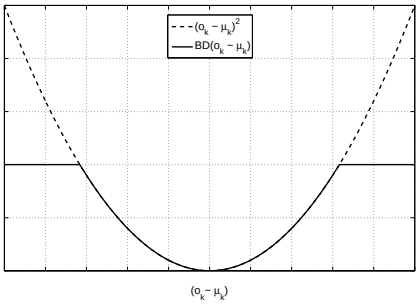

Fig. 1. Bounded euclidean distance (solid line) vs. euclidean distance (dashed line).

dominates the summation if one of the components, o_k, is strongly corrupted and, consequently, is far away from its corresponding mean, μ_k. Thus, the corrupted components contribute strongly towards discarding the corresponding model. For this reason, it is appropriate to bound the distance and, as a result, to bound the influence of the outliers on the final decision. To that purpose, the normalised euclidean distance in eq. (4) is substituted by a bounded distance as follows:

$$-\frac{1}{2}\sum_{k=1}^{N}\frac{BD(o_k-\mu_k)}{\sigma_k^2} \tag{5}$$

where

$$BD(o_k-\mu_k)=\begin{cases}(o_k-\mu_k)^2, \text{ if } |o_k-\mu_k|<\alpha\sigma_k\\ (\alpha\sigma_k)^2, \qquad \text{otherwise}\end{cases} \tag{6}$$

where α is a parameter that controls the actual value of the bound. Figure 1 illustrates the original euclidean distance and the bounded distance used in this work. For clarity reasons, it is worth noting here that for the Gaussian mixture case eq. (6) is applied in every Gaussian component. In addition, it should be noted that the use of the bounded-distance implies that a proper pdf is no longer used.

The underlying idea of BD-HMM is not new. It was first proposed in the scope of speaker recognition (Matsui and Furui (1991, 1992)). Matsui and Furui (1991) use a new distance called DIM (Distortion-Intersection Measure) in a vector quantisation-based speaker recognition system. Later, this distance was adapted to a HMM-based speaker recogniser (Matsui and Furui (1992)). In the latter work, the Gaussian distributions that model the state density probability functions are flattened in order to set a limit to the likelihoods attained from features that lie far away from the mean of the corresponding Gaussians. In particular, the limit is set to 3 times the standard deviation of the Gaussian distribution, which is similar to setting $\alpha = 3$ in eq. (6).

In the field of ASR a comparable proposal was reported under the name of acoustic backing-off (de Veth et al. (1998, 2001a,b)). The authors of these

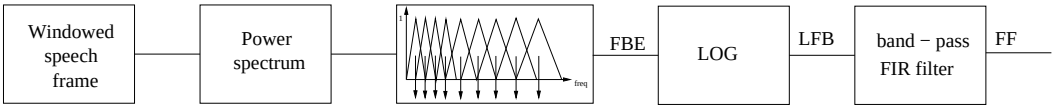

Fig. 2. Frequency filtered parameterisation scheme.

works claim that the classical Gaussian distribution assumed in the HMM framework does not conveniently model the unseen features. They consider that the trained Gaussian distribution proves to be adequate only for features that are well-represented in the training set; however it does not fit well at all the unseen features. Consequently, they propose to use a distribution composed of two weighted terms: the Gaussian distribution estimated in the training phase and an uniform distribution (assuming no previous knowledge) that tries to describe the unobserved features. As a result, the proposed distribution becomes close to a Gaussian which saturates for some determined value (to both sides of the mean) instead of vanishing toward zero.

As long as the distribution for modelling unseen samples in the acoustic backing-off context (de Veth et al. (2001a)) is approximated by a uniform distribution, this method closely resembles the BD-HMM described here. The role of the control parameter α in eq. (6) is now played by the probability value that defines the uniform distribution. However, de Veth et al. (2001a) also change the pdf for the features that are not considered outliers, while BD-HMM does not.

The effectiveness of acoustic backing-off notably depends on the parameterization (de Veth et al. (2001b)). Specifically, it turns out to be more effective for parameterisations, like Frequency Filtered (FF) (Nadeu et al. (1995, 2001); Paliwal (1999)), that do not spread a localized distortion over all the coefficients of the feature vector. Figure 2 summarises by means of a block diagram the steps involved in the FF computation. As can be seen in this figure, 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 involves two log filter bank energies. As a result, the potential localized distortion affecting one log-spectrum feature is just spread over two coefficients of the final parameterisation. In contrast, the DCT involved in the computation of the MFCC parameters spread the distortion over all the coefficients. Taking this conclusion into account,the FF parameterisation is used in this work.

In de Veth et al. (2001b) three real noises were tested in the reported experiments, namely: car, babble and factory noises extracted from NOISEX database (Varga et al. (1992)). The authors found that their method was effective for car noise, a narrow-band and localized noise, but it was not effective for the others. They claim that this is due to the fact that factory and babble noises are wide-band noises while the car noise is clearly colored. As shown

later, the joint utilisation of BD-HMM and spectral subtraction proposed in this paper reaches quite satisfactory results for the three above mentioned noises.

2.2 BD-HMM and missing features

Acoustic backing-off was already interpreted as a Missing Feature (MF) technique (de Veth et al. (2001a)) and BD-HMM can also be interpreted within this framework. MF methods (Cooke et al. (2001); Raj et al. (2004)) focus on the observed features that are considered reliable and, therefore, they discard the use of the unreliable features in the recognition process. Two main approaches can be found in the literature, those that modify the recogniser to admit an incomplete set of features (a subset composed just for the reliable features), and those that estimate the unreliable components before the recognition stage. Both approaches have shown to provide robust solutions as long as an accurate detection of the reliable/unreliable observed features can be obtained. The design of such a classifier is the main problem of these methods and remains an open problem (Raj (2000); Seltzer et al. (2004)). Readers interested in a more comprehensive treatment are referred to Raj and Stern (2005).

Like MF, BD-HMM tackles the recognition problem minimising the influence of the unreliable (outliers) features. However, the outlier (unreliable) or non outlier (reliable) classification is implicitly embedded in the evaluation of the emission probabilities and the design of an explicit classifier is avoided. The BD-HMM specifically focuses on outliers, i. e., remarkably unreliable features, while typical MF approaches aim at making the more difficult reliable/unreliable decision.

3 A synergic combination of bounded-distance HMM and spectral subtraction

3.1 Spectral subtraction

Spectral Subtraction (SS) is a classical speech enhancement method. It consists of removing an estimate of the noise spectrum from the spectrum of the noisy speech signal. Several versions of SS can be found in the literature. In Gong (1995) the interested reader can find a summary of the most relevant proposals.

For this paper we used the minimum statistics method proposed by Martin (2001) to perform the estimation of the noise power spectral density. Essentially, the method looks for minima in the power spectrum of the contaminated speech signal. These minima likely correspond to silence fractions where there is not speech signal and, therefore, those power spectrum minima are related to the mean noise power spectral density through a constant factor. This method allows to update the noisy estimation more often than methods based on voice activity detectors (VADs), since it works even during the short silence periods between words or sentences, and, as a result, is more appropriate for non-stationary noises.

In this work, we have taken ideas from different authors to implement a particular SS method. Specifically, our implementation is based on the following ideas:

- Before removing the noise power spectral estimate from the contaminated power spectrum, Boll (1979) suggests to make an average of this latter spectrum over time. This average should involve just a few frames due to the inherent non-stationary nature of the speech signal. We used three frames (30 ms) for computing this average. Assuming that the noisy signal is stationary along these three frames we are reducing the variance of the noise power spectral density by a factor of three.
- It is more convenient to perform SS at the input of the mel-scaled filter bank (Nolazco Flores and Young (1994)). The mel-scaled filter bank involves an averaging operation within each one of the considered bands. Thus, it is preferable to carry out the estimation of the noise spectrum and the SS before this averaging. On the contrary, some spectral resolution would be lost.
- When the noise power spectrum estimation is subtracted from the noisy speech power spectrum, the result could become negative as a consequence of potential estimation errors. Several "ad-hoc" solutions have been proposed to solve this problem (Gong (1995)). Among them, we have chosen one which establishes a minimum level for the power spectrum of the enhanced speech. This minimum level is given by the noise power spectrum attenuated by a constant factor. Such an scheme is commonly used and a recent example can be found in Pujol et al. (2004).

To sum up, the SS method implemented in this paper can be summarised by the next equation:

$$\widehat{P_X}(\omega, l) = \max\left\{ \left(\widehat{P_S}(\omega, l) - \gamma\widehat{P_N}(\omega, l)\right), \beta\widehat{P_N}(\omega, l)\right\} \tag{7}$$

where $\widehat{P_X}$, $\widehat{P_S}$ and $\widehat{P_N}$ are, respectively, the estimates of the clean, noisy speech and noise power spectrum densities; 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". As indicated above, $\widehat{P_S}$ is estimated as an average over the power spectrum densities at the previous, current and next time frames while $\widehat{P_N}$ is estimated using the method proposed by Martin (2001).

It is well-known that, although SS methods can improve the SNR, the nonlinear operations involved produce distortions that may degrade the ASR system performance (Gong (1995)). In the next section we will see how bounded-distance HMM can be an excellent companion method for circumventing the negative effects of these nonlinear distortions on the performance of ASR systems.

3.2 A synergic combination of bounded-distance HMM and spectral subtraction

The BD-HMM method is effective to cope with outliers that appear when ASR systems deal with noisy speech. As explained in Section 2 the BD-HMM method limits the normalized euclidean distance in the Gaussian exponent to a maximum value and, as a result, the negative effects of the outliers are reduced. 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 we will experimentally show in Section 4, these distortions produce an increment of the number of outliers. Taking this fact into consideration, in this paper we propose the joint application of SS and BD-HMM. To be more precise, we suggest to take 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, that compensates all the parameters (not just the outliers).

3.3 Implementation details

The implementation of the two considered methods in the proposed combination offers some degrees of freedom that deserve to be discussed and clarified before presenting our experimental results.

For the BD-HMM method we only need to set the parameter α in eq. (6) that determines which features are outliers. In our implementation a value $\alpha = 3$ was employed that also coincides with the value used by Matsui and Furui (1992). The limit imposed by this value for α does not degrade the performance of the system in absence of outliers, since, as it is well known, for a Gaussian distribution the 99.7 % of the samples fall within $\pm 3\sigma$ (the experimental results that show the good performance of the method in clean conditions are given in Table 2, Section 4).

With respect to the free parameters of the SS method, γ and β in (eq. (7)), we explored a small set of values, namely: $\gamma = \{0.8,\ 1.0\}$ and $\beta = \{0.1,\ 0.2,\ 0.3\}$ for every SNR and noise type for the RM1 task (this task will be described in detail in Section 4). Although none set of values achieved the best results under all conditions, the set $\{\gamma = 0.8,\ \beta = 0.2\}$ was the most suitable when SS was used alone. On the other hand, when SS was used in combination with BD-HMM, the preferred set of parameters was $\{\gamma = 1.0,\ \beta = 0.1\}$. The complete set of results that allowed us to infer this last conclusion is shown later in Subsection 4.4.1. These parameter values for the task RM1 were extrapolated to the rest of the tasks that will be also described in Section 4.

4 Experiments and Results

Two types of experiments were used to assess the proposed method. In the first experiments, in order to motivate our proposal, we measured the influence of the outliers on the recognition process. In the second experiments we obtained results in terms of the word error rate for several ASR tasks that support our proposal.

Four different noisy speech recognition tasks were used to perform the experiments. In the following subsection we present the setting that was shared by the four tasks. Next, we describe each task in more detail. Finally, the results for the two types of experiments are presented.

4.1 System set-up

As mentioned above, four different tasks were considered in our experiments. Each one was set up using a well-known database, namely: RM1 (NIST (1992)), Wall-Street Journal (Paul and Baker (1992)), Aurora-4 (Hirsch (2002)) and Spanish SDC-Aurora (Macho (2000)). For each task we built a specific baseline ASR system using the HTK toolkit (Young et al. (2002)).

The same parameterisation was used for all experiments. Specifically, the feature vector consists of 12 FF parameters plus the log-energy coefficient that were obtained every 10 ms using a 25 ms Hamming window. For extracting the 12^{nd}-dimensional FF vector we computed 14 log filter bank energies and we ignored the first and last samples of the filtered sequence. The time mean of the FF parameters along each utterance was removed, the log-energy was normalised and the resulting feature vector was extended with the first and second time derivatives, resulting in a 39-dimensional vector. The training set was also processed using SS as proposed by Shozakai et al. (1997).

4.2 Database descriptions

4.2.1 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 an unique state tied to the middle state of the long silence model. Finally, the standard word-pair grammar was used as the language model.

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

4.2.2 Wall Street Journal (WSJ0) (Paul and Baker (1992))

The Wall Street Journal (WSJ0) database was the second database considered. The standard SI-84 training set, which contains 7138 utterances, was used to build the models. For evaluation, we employed the Nov.'92 CSR Speaker-Independent 5K Read NVP (Non Verbalisation Punctuation) test set. Again, an 8 kHz down-sampled version of the database was used.

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 triphones acoustic models and two different silence models were used. In this case, three-state models with 8 Gaussian per state were used, except for the silence models in which the number of Gaussians was increased until 16. The 5K bigram language model distributed with the corpus was used.

Finally, the artificially created noise-contaminated versions for evaluation purposes were the same as for the RM1 database case.

4.2.3 Aurora-4 (Hirsch (2002))

This database is based on the above WSJ0 database and the Nov'92 WSJ0 development test.

As training set, we chose the predefined set, which comprises the clean speech sentences acquired with the close-talking microphone. This set agrees with the training set used for the design of the WSJ0 system described in the previous section. The test set was based on the Nov'92 development set 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 noise 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 baseline ASR system was the same as for the WSJ0 database.

4.2.4 Spanish SDC-Aurora (Macho (2000))

Unlike the previously considered databases, in the Spanish SDC-Aurora the noise was not artificially added to the clean speech, but the speech was directly recorded in a noisy environment.

Spanish SDC-Aurora database comprises 4914 recordings using both a close-talking microphone and a hands-free microphone. The recordings were made in three different noisy conditions: quiet (inside a car, stopped engine), low (driving at low speed on a town road) and high (driving at high speech in a good road). The sentences were acquired at 16 kHz and subsequently transformed to 8 kHz. The Well-Matched (WM), Medium-Mismatch (MM) and High-Mismatch (HM) standard experiments for this database were considered. For the description of such a standard experiments the reader is referred to Macho (2000).

Table 1
Summary of the main differences between the 4 considered ASR tasks

	RM1	**WSJ0**	**Aurora-4**	**Spanish SDC-Aurora**
Number of words	991	5000	5000	10
Distortions types	Contaminated for our experiments: 5 noises and 4 SNRs	Contaminated for our experiments: 5 noises and 4 SNRs	6 noises at a randomly chosen SNR	1 real noise, 3 noisy conditions

Isolated digit and connected digits experiments were carried out. The baseline ASR system for this task was built using the scripts distributed with the database. Eighteen-state digit models were built using 3 Gaussians per state. As for the previous tasks, two silence models were distinguished, for modelling either long or short pauses between digits. The long one used 3 states and 6 Gaussians per state while the short used only one state that was tied to the middle state of the long pause model.

Finally, Table 1 provides a brief summary of the main differences between the 4 considered tasks.

4.3 On the influence of spectral subtraction on outliers

First, we measured the percentage of outliers that occurs in the recognition process when the speech signal is contaminated by additive noises. These experiments were carried out for the test set of the RM1 database in two cases: with and without SS. Since the acoustic models are the ones that decide if an observed feature is an outlier or not, the percentage of the outliers depends on the recognition path. For this reason, the experiments were conducted using the state sequence that was obtained by means of a forced alignment between the correct labels and the parameters extracted from the clean speech.

Figure 3 shows the results. The white bars indicate the percentage of outliers computed directly from the original noisy parameters, while the black bars indicate the additional percentage of outliers computed once SS was applied (in this experiment we have used the parameter set $\{\gamma = 1.0, \beta = 0.1\}$ in eq. (7)). As can be observed from the figure, the number of outliers when SS is applied increases significantly and consistently. As expected, the number of outliers increases as the SNR decreases. It is also worth noting that these percentages are low; however, as we show in the next experiment, their influence on the

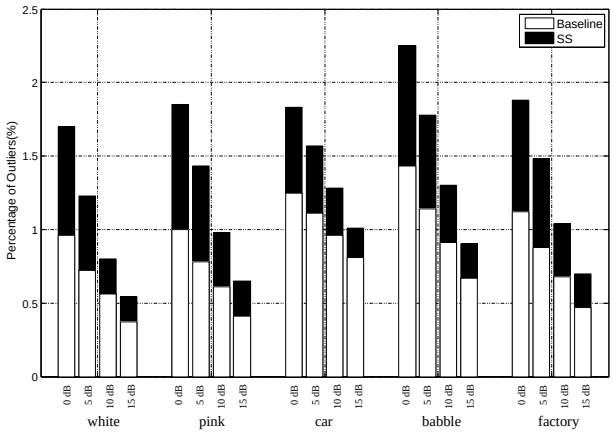

Fig. 3. Percentage of outliers found in the RM1 database test set for several noise types and SNRs. The bars labelled as *Baseline* refer to the percentage computed from the original noisy parameters, while the bars labelled as *SS* show the percentage obtained when spectral subtraction was applied to compensate the additive distortion.

recogniser decision is quite significant.

When an unknown sentence is recognised, the log-likelihood along all the possible paths is computed and the maximum is chosen. We now consider the part of the log-likelihood that accounts for how well the observations are modelled by the HMM emission probability distributions. Going back to eq. (1), we are referring to the time-accumulated emission log-probabilities, i.e.:

$$\log(b)_{accum} = \sum_{t=1}^{L} \log(b^i_{x_t}(\mathbf{o}_t)) \tag{8}$$

In order to evaluate the potential influence of the outliers on the recogniser decision, we measured the outlier contribution to this time-accumulated log-probability. Again, since this time-accumulated log-probability depends on the considered recognition path, we conducted our experiment using a forced alignment. In this context, the term $\log(b)_{accum}$ was computed in two ways: with and without BD-HMM. Let $\log(b)_{accum}$ denote the computed value without using BD-HMM and let $\log(b)^{BD-HMM}_{accum}$ denote the result obtained using BD-HMM. Finally, in order to quantify the contribution of the outliers to the log-likelihood the following percentage was computed for every sentence in the RM1 database test set:

$$D(\%) = 100 \frac{\log(b)^{BD-HMM}_{accum} - \log(b)_{accum}}{\sum_{t=1}^{L} |\log(b^i_{x_t}(\mathbf{o}_t))|} \tag{9}$$

In order to make the experiment more intuitive and clearer, the BD-HMM method was modified emphasising the concept of outlier. Specifically, the

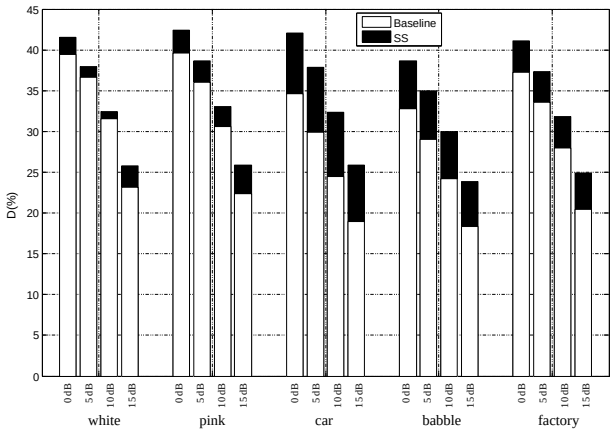

Fig. 4. Average percentage contribution of the outliers to the accumulated emission log-probability on the RM1 database test set for several noise types and SNRs. The bars labelled as *Baseline* refer to the term computed from the original noisy parameters, while the bars labelled *SS* show the percentage obtained when spectral subtraction was applied to compensate the additive distortion.

bounded normalised euclidean distance (eq. (5)) was only applied when the feature was out of the scope of *all* the Gaussians that make up the Gaussian mixture. In the original form of the BD-HMM method the bounded distance is independently applied to every Gaussian. Thus, every Gaussian takes the outlier/non outlier decision. If a feature is an outlier for *all* the Gaussians it will be also an outlier for *every* Gaussian. As a result, the percentage D that we computed is a subset of the percentage detected in the original BD-HMM method.

Figure 4 shows the mean value of the percentage contribution of the outliers to the accumulated emission log-probability considering all the sentences in the RM1 database test set, for several noise types that were artificially added to the original speech data at several SNRs. Furthermore, we present the results with (black bars) and without SS (white bars). From this figure, the significance of the outliers in the recognition process becomes revealing. In general, they explain between the 20 % and the 40 % of accumulated log-probability. In addition, the mean value of the percentage contribution of the outliers significantly increases, as expected, due to the use of SS. It is evident that, even though the outliers do not contain any relevant information concerning the embedded message, they have a significant weight on the recogniser decision. Therefore, the role of BD-HMM is justified.

4.4 Experimental assessment of the proposed method

The suggested combination of SS and BD-HMM (Henceforth, *SSBD-HMM*) was assessed for the four previously described ASR tasks. Performance results

Table 2
WER (%) for clean speech attained by each of the compared systems for the RM1, WSJ0 and Aurora-4 tasks.

	Baseline	**SS**	**BD-HMM**	**SSBD-HMM**
RM1	6.70 %	6.63 %	6.29 %	6.59 %
WSJ0	9.88 %	9.79 %	9.02 %	9.30 %
Aurora-4	9.38 %	9.79 %	8.84 %	9.25 %

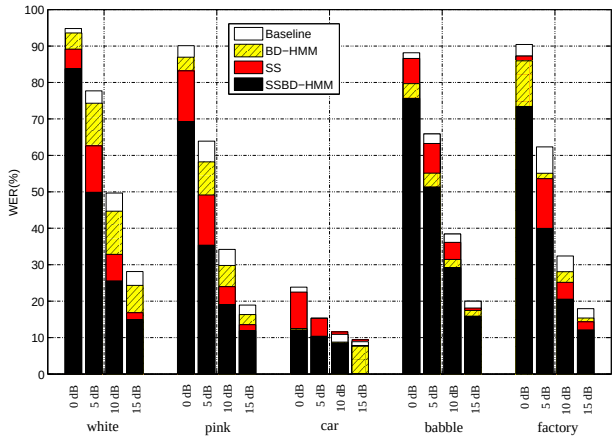

Fig. 5. WER (%) attained by each of the compared systems for the RM1 task.

in terms of Word Error Rate (WER) are given along with those attained by three reference systems, namely: the baseline (*Baseline*), the baseline plus SS (*SS*) and the baseline plus BD (*BD-HMM*). Moreover, a more detailed analysis is presented only for the RM1 case.

Before presenting and discussing the results for every task, recognition results in terms of WER for clean speech are shown in Table 2 for reference purpose (results for the Spanish SDC-Aurora task are not shown since the clean speech is not available). It is observed that, although BD-HMM gets slightly better results for clean speech, none of the methods produce significant changes in the recognition rates achieved by the baseline system.

4.4.1 Results and supplementary analysis for RM1

The WER for all the compared systems and several noise types and SNRs are given in Figure 5.

As can be observed, the SSBD-HMM method clearly outperforms the reference systems in all of the cases with the only exception of the car noise at 15 dB, for which the BD-HMM is slightly superior to the combination SSBD-HMM.

For white, pink and factory noises SS achieves superior performance to BD-

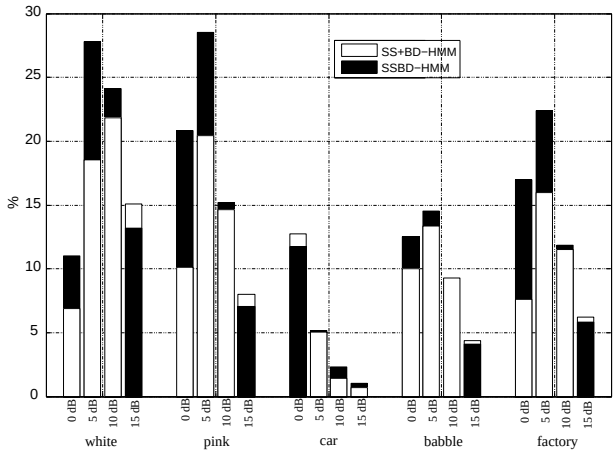

Fig. 6. WER reduction obtained by the combination SSBD-HMM compared to the sum of the WER reductions individually achieved by SS and BD-HMM, labelled as *SS+BD-HMM*.

HMM (with the only exception of factory noise at 0 dB). For babble and car noises the reverse situation is found: SS gets poorer results than BD-HMM. Car noise is the one for which the acoustic backing-off method (de Veth et al. (2001b)) achieved its best performance; in our experiments we observe the same trend, BD-HMM attains notable improvements by itself and no significant advantage is derived from the SSBD-HMM combination. On the other hand, the use of SS does not achieve any improvement by itself, even incurring in some performance losses for the highest SNR; however, these losses are compensated by BD-HMM when using SSBD-HMM.

It is also interesting to highlight the synergy found between both methods, SS and BD-HMM, in most cases. This synergy becomes evident when the combination SSBD-HMM achieves a greater performance improvement than the sum of those due to either SS or BD-HMM. This fact is illustrated in Figure 6, which shows both the WER reduction achieved by SSBD-HMM and the sum of the WER reductions individually achieved by SS and BD-HMM, which is labelled as *SS+BD-HMM*.

As observed in this figure, the synergy becomes clear for lower SNRs (except for car noise) and vanishes as the SNR is higher. These results agrees with the fact that the SS method produces a higher number of outliers, whose effect is alleviated by BD-HMM, when the SNR is lower.

As previously mentioned (Subsection 3.3), the chosen values for the parameters γ and β are different if SS is considered alone or in combination with BD-HMM. Figure 7 shows the attained results, in terms or WER, by both SS and SSBD-HMM for two sets of parameters, namely: $\{\gamma = 0.8,\ \beta = 0.2\}$ and $\{\gamma = 1.0,\ \beta = 0.1\}$.

In most of the cases, the set $\{\gamma = 0.8,\ \beta = 0.2\}$ is optimal for SS. However,

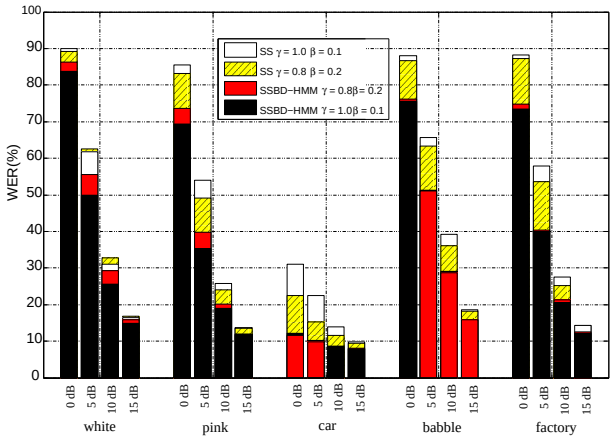

Fig. 7. Performance dependency of SS and SSBD-HMM on SS parameters γ and β. WER (%) achieved by SS and SSBD-HMM for two sets of parameters, $\{(\gamma = 0.8, \beta = 0.2), (\gamma = 1.0, \beta = 0.1)\}$, on the RM1 task.

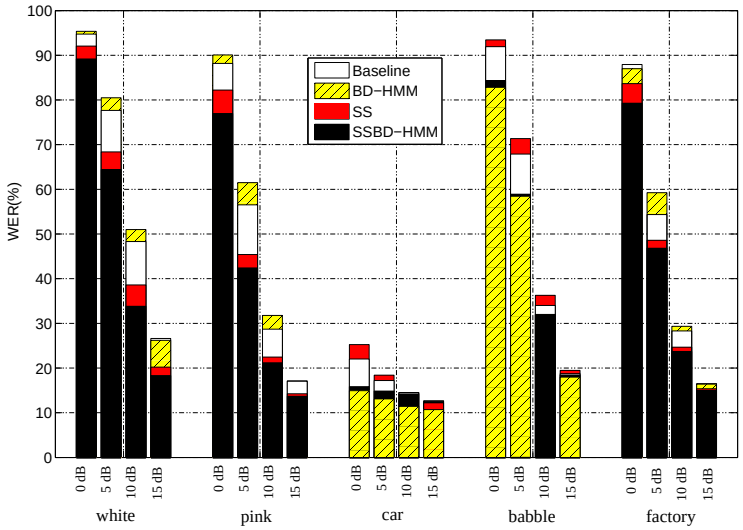

Fig. 8. WER (%) attained by each of the compared systems for the WSJ0 task.

for SSBD-HMM, the best pair is $\{\gamma = 1.0,\ \beta = 0.1\}$. This fact is accounted for the different number of outliers generated by each set of parameters. The set $\{\gamma = 1.0,\ \beta = 0.1\}$ generates more outliers and, therefore, SS gets poorer results. However, when BD-HMM is also applied these outliers do not have influence on the decision process and better results are achieved. In order words, the set $\{\gamma = 1.0, \beta = 0.1\}$ is better for enhancing the speech parameters at the expense of introducing a larger number of outliers.

4.4.2 Results for WSJ0, Aurora-4 and Spanish SDC-Aurora

Experiments under similar conditions were conducted for the three remaining tasks, namely: WSJ0, Aurora-4 and Spanish SDC-Aurora.

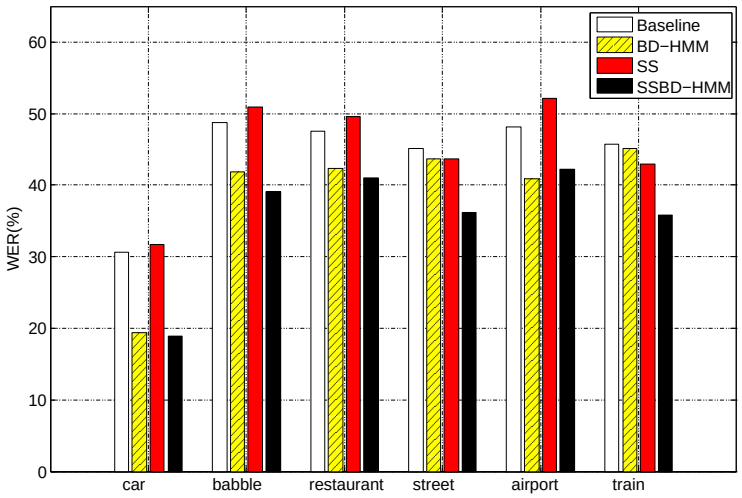

Fig. 9. WER (%) attained by each of the compared systems for the Aurora-4 task.

Figure 8 shows the results for WSJ0. The suggested combination SSBD-HMM again achieves the best performance for most of the noise types and SNRs. The exceptions occur for car and babble noises, for which BD-HMM alone becomes the best choice, while SS does not work. Nevertheless, in these cases, the performances of BD-HMM and SSBD-HMM are similar, i.e., BD-HMM is able to compensate for the poor SS performance. In these cases, the SSBD-HMM method obtained better results with the set $\{\gamma = 0.8,\ \beta = 0.2\}$ while Figure 8 shows the results for the set $\{\gamma = 1.0,\ \beta = 0.1\}$. For factory noise at medium SNRs the situation is even more favourable to our proposal: SS does not work but the combination SSBD-HMM outperforms BD-HMM. For white and pink noises BD-HMM does not perform well, in contrast to SS. The combination SSBD-HMM clearly outperforms SS, making evident the claimed synergy.

Figure 9 displays the results for the Aurora-4 task. The combination SSBD-HMM again gets the best results in most of the cases. As seen in the figure, SS does not result in any improvement (with one exception). Nevertheless, the synergy becomes apparent, since the combination SSBD-HMM clearly outperforms BD-HMM in almost all the cases.

For car and babble noises the results are similar to those obtained for the WSJ0 task. It is worth noting that a slightly better performance is achieved for babble noise, likely due to the different SNR used for contaminating the sentences.

Finally, we used the Spanish SDC-Aurora database to asses our proposal with non-simulated noise addition (see Subsection 4.2.4 for details). Figure 10 shows the results for the three standard experiments defined in the database. Once more the combination SSBD-HMM takes the best of both methods. The claimed synergy is remarkable in the High-Mismatch (HM) experiment, in

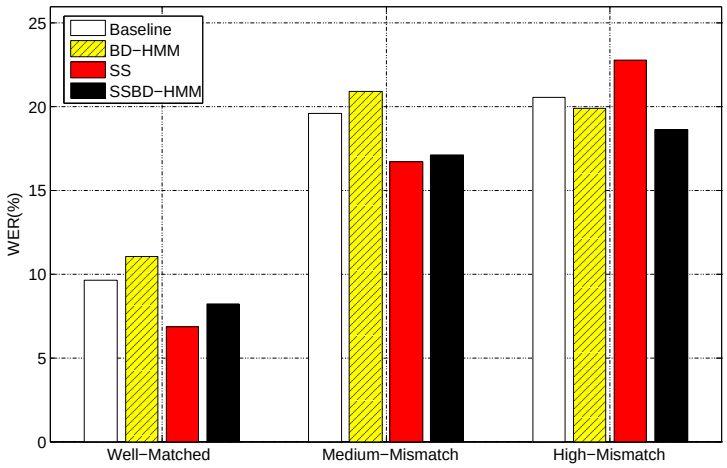

Fig. 10. WER (%) attained by each of the compared systems for the Spanish SDC-Aurora task.

which BD-HMM achieves a slight improvement with respect to the baseline, SS causes a notable loss of performance and the combination gets a significant improvement (clearly superior to that achieved by BD-HMM).

5 Conclusions

In this paper we propose the combination of what we call bounded-distance HMM and spectral subtraction as an effective method to deal with additive noise in ASR.

The BD-HMM method aims at mitigating the effect of outliers on the recogniser decision. In this paper the use of BD-HMM is motivated by quantifying the effect of the outliers on the log-likelihood which determines the ASR system decision. We note that BD-HMM is similar to acoustic backing off (de Veth et al. (2001a)), though the implementation differs. Our experimental results allow us to conclude that the combination method suggested in this paper overcomes some of the limitations reported in de Veth et al. (2001a).

BD-HMM is limited in the sense that it acts only on the outliers. To overcome this limitation, the combination of BD-HMM and SS is suggested. Our results allow us to conclude that SS is an excellent companion method. Specifically, we show how the use of SS significantly increases the number of outliers, which affects ASR systems performance. BD-HMM compensates for this side-effect of SS. Thus, it is possible to benefit from the best aspects of both methods simultaneously: the speech feature enhancement due to additive noise removal associated with SS and the ability of BD-HMM to reduce the impact of signal distortion on the recognition process. A clear synergy between both methods is obtained and the combination outperforms the sum of the improvements of

the individual methods considered separately.

Acknowledgements

This work has been partially supported by Spanish Regional grant CCG06-UC3M/TIC-0812.

References

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.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., Jun. 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34 (3), 267–285.

de la Torre, A., Peinado, A., Segura, J., Perez-Cordoba, J., Benitez, M., Rubio, A., May 2005. Histogram equalization of speech representation for robust speech recognition. IEEE Trans. Speech Audio Processing 13 (3), 355–366.

de Veth, J., Cranen, B., Boves, L., Dec. 1998. Acoustic backing-off in the local distance computation for robust automatic speech recognition. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP). pp. 1427–1430.

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., Acero, A., Plumpe, M., X.Huang, Oct. 2000. Large-vocabulary speech recognition under adverse acoustic environments. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP). pp. 806–809.

Furui, S., Apr. 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoustics, Speech, Signal Processing 29 (2), 254–272.

Gales, M., Woodland, P., 1996. Mean and Variance Adaptation within the MLLR framework. Computer Speech & Language 10 (4), 249–264.

Gales, M., Young, S., Sep. 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Processing 4 (5), 352–359.

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

Hain, T., Woodland, P., Niesler, T., Whittaker, E., Mar. 1999. The 1998 HTK system for transcription of conversational telephone speech. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP). Vol. 1. pp. 57–60.

Hermansky, H., Morgan, N., Oct. 1994. RASTA processing of speech. IEEE Trans. Speech Audio Processing 2 (4), 578–589.

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.

Macho, D., Nov. 2000. Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: Description and baseline results. Tech. rep., UPC, Universitat Politecnica de Catalunya.

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.

Matsui, T., Furui, S., Apr. 1991. A text-independent speaker recognition method robust against utterance variations. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP). Vol. 1. pp. 377–380.

Matsui, T., Furui, S., Mar. 1992. Comparison of test-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP). Vol. 2. pp. 157–160.

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.

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.

Nolazco Flores, J., Young, S., Apr. 1994. Continuous speech recognition in noise using spectral subtraction and HMM adaptation. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP). Vol. 1. pp. 409–412.

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.

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.

Pujol, P., Nadeu, C., Macho, D., Padrell, J., Sep. 2004. Speech recognition experiments with the SPEECON database using several robust front-ends. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP).

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

Raj, B., Apr. 2000. Reconstruction of incomplete spectrograms for robust speech recognition. Ph.D. thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Raj, B., Seltzer, M., Stern, R., Sep. 2004. Reconstruction of missing features for robust speech recognition. Speech Communication 43 (4), 275–296.

Raj, B., Stern, R., Sep. 2005. Missing-feature approaches in speech recognition. IEEE Signal Processing Magazine 22 (5), 101–116.

Seltzer, M., Raj, B., Stern, R., Sep. 2004. A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Communication 43 (4), 379–393.

Shozakai, M., Nakamura, S., Shikano, K., Dec. 1997. A speech enhancement approach E-CMN/CSS for speech recognition in car environments. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 450–457.

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.

Viikki, O., Laurila, K., Aug. 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Communication 25 (1), 133–147.

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.