State-based labelling for a sparse representation of speech and its application to robust speech recognition

Tuomas Virtanen[1]*, Jort F. Gemmeke*[2]*, Antti Hurmalainen*[1]

[1]Department of Signal Processing, Tampere University of Technology, Finland,
[2]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
tuomas.virtanen@tut.fi, j.gemmeke@let.ru.nl, antti.hurmalainen@tut.fi

Abstract

This paper proposes a state-based labeling for acoustic patterns of speech and a method for using this labelling in noise-robust automatic speech recognition. Acoustic time-frequency segments of speech, exemplars, are obtained from a training database and associated with time-varying state labels using the transcriptions. In the recognition phase, noisy speech is modeled by a sparse linear combination of noise and speech exemplars. The likelihoods of states are obtained by linear combination of the exemplar weights, which can then be used to estimate the most likely state transition path. The proposed method was tested in the connected digit recognition task with noisy speech material from the Aurora-2 database where it is shown to produce better results than the existing histogram-based labeling method.

Index Terms: noise robustness, automatic speech recognition, sparse representations

1. Introduction

Most of automatic speech recognition (ASR) technologies are based on hidden Markov models (HMMs), which model a time-varying speech signal using a sequence of states, each of which is associated with a distribution of acoustic features. While HMMs reach a relatively high performance in good conditions, they have problems in modelling wide variances in natural speech signals, such as speech in natural environments which is often interfered by environmental noises.

Recently, some studies [1–6] have aimed at ASR using sparse representations of speech. In them, a time-frequency representation of speech is as a weighted linear combination of speech atoms. Benefits of the existing systems range from improved recognition accuracy to an easy incorporation of robustness to additive noises. Some of these systems construct the dictionary of atoms to be used in the sparse representation from exemplars of speech, which are realizations of speech in the training data, spanning multiple time frames [1].

When the weights of the sparse representation are used directly in the recognition, a fundamental problem is the association of higher-level information with the atoms in the dictionary to enable the recognition. Sivaram et al. [3] trained a neural network to map the weights of the atoms directly to phoneme classes. Sainath et al. [4] associated each atom with one phonetic class, and recognition was done by finding the phoneme class with the highest sum of weights. Van hamme et al. [5] used a dictionary consisting of both acoustic information and higher-level phonetic information. Schullet et al. [6] used the index of the speech atom with the highest weight as an additional feature for their Dynamic Bayesian Network recognizer. In our earlier work [1] we did exemplar-based ASR, where each exemplar was associated with a state histogram which expressed the occurrence count for each HMM-state within the duration of the exemplar. With the state transcription of the dictionary obtained by forced alignment, likelihoods were then calculated as the weighted sum of exemplar likelihoods.

With exemplars spanning multiple frames, it becomes increasingly difficult to accurately model the time-varying information. For example the histogram-based representation [1] does not carry information about ordering of the states within an exemplar, and it is unable to represent fine-level temporal information. This can lead to erroneous recognition, especially in the case of repeated phonemes and in short utterances.

This paper proposes an exemplar-based ASR system where the higher-level information about each exemplar is encoded with a state label matrix. The state label matrix of each exemplar represents the state activity of the exemplar as a function of time. The state activities are calculated separately for overlapping segments of speech, and the state likelihoods for a whole utterance are obtained as a weighted sum of overlapping exemplars. We also propose a new method to obtain the likelihoods of silence states by estimating the speech activity from the exemplar weights. We evaluate the proposed method on the AURORA-2 noisy connected digit recognition task, where it is shown to produce better results than the existing histogram-based labeling method.

2. Sparse representation of noisy speech

The proposed approach operates in the magnitude spectrogram domain, with the term magnitude spectrogram referring to the square root of energy as the function of time and frequency. The magnitude spectrogram describing a clean speech segment $\mathbf{S}$ is a $B \times T$ dimensional matrix (with B frequency bands and T time frames). To simplify the notation, the columns of this matrix are stacked into a single vector $\mathbf{s}$ of length $D = B \cdot T$, so that the entry $\mathbf{S}(b,t)$ corresponds to the entry $\mathbf{s}(b+tB)$. Top panels of Figure 1 illustrate two examples of exemplars.

We model an observed speech segment as a linear, non-negative combination of clean speech exemplars $\mathbf{a}^{\mathrm{s}}_j$, $j = 1,\ldots,J$ denoting the exemplar index. These stacked vectors are magnitude spectrograms describing speech segments extracted from a training database. We write:

$$\mathbf{s} \approx \sum_{j=1}^{J} x^{\mathrm{s}}_j \mathbf{a}^{\mathrm{s}}_{\mathbf{j}} = \mathbf{A}^{\mathrm{s}}\mathbf{x}^{\mathrm{s}} \quad \text{s.t.} \quad \mathbf{x}^{\mathrm{s}} \geq 0 \tag{1}$$

where $\mathbf{x}^{\mathrm{s}}$ is a J-dimensional weight vector and the J exemplars $\mathbf{a}^{\mathrm{s}}_1 \ \mathbf{a}^{\mathrm{s}}_2 \ldots \mathbf{a}^{\mathrm{s}}_J$ are grouped into a speech exemplar matrix $\mathbf{A}^{\mathrm{s}}$ as $\mathbf{A}^{\mathrm{s}} = [\mathbf{a}^{\mathrm{s}}_1 \ \mathbf{a}^{\mathrm{s}}_2 \ldots \mathbf{a}^{\mathrm{s}}_J]$. Prior research has shown [7] that $\mathbf{x}^{\mathrm{s}}$ can be extremely *sparse*. That is, only a few nonzero entries suffice to represent $\mathbf{s}$ with sufficient accuracy. The weights are restricted to non-negative values, a restriction which has turned

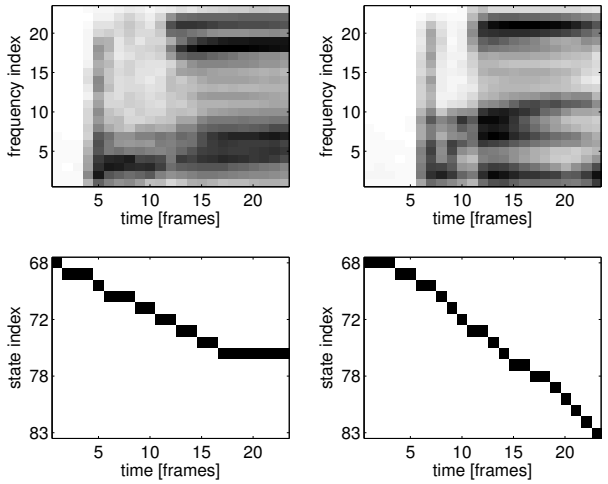


Figure 1: Two exemplars of length $T = 25$ representing two different realizations of the digit "two". The horizontal axes indicate the time in frames. The top panels illustrate the magnitude spectra, with a dark color indicating a higher value. The lower panels illustrate a part of the state label matrices explained in Section 3.1. The state indices $[68, 83]$ are the 16 states underlying the digit "two". The digit in the left panels has a longer duration, and therefore it does not reach the higher state indices that the right exemplar does. The histogram-based method [1] uses the state occurrence histograms of these label matrices.

out to be very critical in audio analysis algorithms based on the linear model for the magnitude spectrogram [8].

Noisy speech signals are modelled by including a dictionary $\mathbf{A}^{\mathrm{n}}$ which contains noise atoms, so that we get

$$\mathbf{s} \approx [\mathbf{A}^{\mathrm{s}}\mathbf{A}^{\mathrm{n}}]\,[\mathbf{x}^{\mathrm{s}}\mathbf{x}^{\mathrm{n}}] \tag{2}$$

$$= \mathbf{A}\mathbf{x} \quad \text{s.t.} \quad \mathbf{x}^{\mathrm{s}}, \mathbf{x}^{\mathrm{n}}, \mathbf{x} \geq 0 \tag{3}$$

where $\mathbf{x}^{\mathrm{n}}$ is the K-dimensional weight vector of the noise exemplars and the whole speech + noise exemplar dictionary matrix $\mathbf{A}$ has dimensionality $D \times L$, where $L = J + K$, and vector $\mathbf{x}$ consists of the weights of the speech and noise exemplars.

2.1. Finding the weights

In order to obtain $\mathbf{x}$, we look for values which are able to represent the noisy speech $\mathbf{y}$ with the model $\mathbf{A}\mathbf{x}$, while using only a small number of non-zero entries in $\mathbf{x}$. We do this by minimizing the cost function

$$d(\mathbf{y}, \mathbf{A}\mathbf{x}) + ||\boldsymbol{\lambda}. * \mathbf{x}||_1 \quad \text{s.t.,} \quad \mathbf{x} \geq 0, \tag{4}$$

where the first term measures the mismatch between the noisy observation and the model by the generalized Kullback-Leibler (KL) divergence:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{d=1}^{D} y_d \log(\frac{y_d}{\hat{y}_d}) - y_d + \hat{y}_d, \tag{5}$$

which has been found out to produce good results in audio spectrogram modeling [8].

The second term of (4) is the L1 norm of the weight vector weighted by element-wise multiplication (operator $.*$) by vector $\boldsymbol{\lambda} = [\lambda_1 \ \lambda_2 \ldots \lambda_L]^{\mathrm{T}}$, which leads to cost $\sum_{l=1}^{L} x_l \lambda_l$. The term penalizes non-zero entries of $\mathbf{x}$, controlling the degree of sparseness of the resulting $\mathbf{x}$. The cost function (4) is minimized using the multiplicative updates routine described in [1].

2.2. Sliding window approach for time-continuity

Describing (noisy) speech as a linear combination of exemplars is only a valid approach for relative short speech segments. In order to decode utterances of arbitrary lengths, we adopt a sliding window approach as in [1]. In this approach, we divide an utterance in a number of overlapping, fixed-length segments. We then find a sparse representation for each segment.

Consider a noisy speech utterance $\mathbf{Y}_{\text{utt}}$ represented as a magnitude spectrogram of size $B \times T_{\text{utt}}$. We slide a window (observed speech segment) $\mathbf{Y}$, a matrix of size $B \times T$, through $\mathbf{Y}_{\text{utt}}$ using a window shift of one frame.

At each window position, the spectrogram is reshaped into an observation vector $\mathbf{y}_w$, similarly as was done for speech and noise exemplars above. Here w denotes the index of the window position, and it ranges from 1 to $W = T_{\text{utt}} - T + 1$, the number of windows in the utterance. For each $\mathbf{y}_w$, we obtain the sparse representation weight vector $\mathbf{x}_w$ as described above.

3. Classification using state-labeling

In this section we propose a hybrid exemplar-based/HMM method for recognizing the words in the observed utterance, based on the speech exemplar weights obtained in the previous section, and a state label matrix associated with each exemplar. The states in our system have a similar role as in conventional HMM recognizers, but we use a exemplar-based method for estimating the likelihoods of the states.

3.1. Speech state likelihoods

Each frame $t = 1, \ldots, T$ in each speech exemplar $\mathbf{a}_j^{\mathrm{s}}$ is labelled with a state label $q_{jt} \in [1, Q]$, where Q is the total number of states. Given the canonical transcriptions of the training data from which the exemplars are extracted, the labelling is obtained by using a forced alignment with a conventional HMM recognizer. The training data used to obtain the labelling is presented in Section 4.

Using the frame-by-frame state labelling of the exemplars, we encode the labelling of each exemplar $\mathbf{a}_j^{\mathrm{s}}$ with label matrix $\mathcal{L}_j$. $\mathcal{L}_j$ is a sparse, binary matrix of dimensions $Q \times T$, the entries having values $[\mathcal{L}_j]_{q,t} = \delta(q, q_{jt})$, where δ is the delta function. Figure 1 illustrates two examples of exemplars and their corresponding state label matrices.

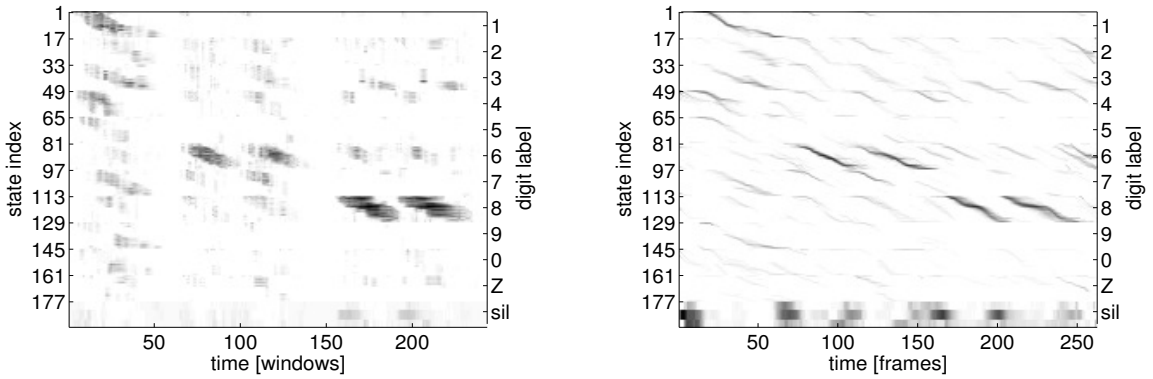
Denoting the speech exemplar weights calculated for window w by $x_{w,j}^{\mathrm{s}}$, $j = 1, \ldots, J$, j being the exemplar index, we calculate state likelihood matrix $\mathbf{L}_w$ in window w as the weighted sum of exemplar label matrices as

$$\mathbf{L}_w = \sum_{j=1}^{J} \mathcal{L}_j x_{w,j}^{s} \tag{6}$$

The columns of $\mathbf{L}_w$ are denoted with vectors $\mathbf{l}_{w,\tau}$, $t = 1, \ldots, T$. State likelihood vectors combined over overlapping windows are obtained by summing the likelihoods of the frames of all the windows that overlap, taking into account the exact temporal positions of the frames. The combined state likelihood vector $\mathbf{l}_\tau^{\text{utt}}$ for each frame $\tau = 1, \ldots, T_{\text{utt}}$ is given as

$$\mathbf{l}_\tau^{\text{utt}} = \sum_{\tau=\max(1,\tau-T_{\text{utt}}+T)}^{\min(T,\tau)} \mathbf{l}_{\tau-t+1,t} \tag{7}$$

Figure 2 illustrates an example of a likelihood matrix where the frame likelihood vectors are the column vectors. In our previous approach [1] we used histograms of state labels instead of dynamic state label matrices. The method proposed in this paper produces sharp likelihoods paths along with the likelihood



(a) Histogram based labelling

(b) Time-dependent labelling

Figure 2: Two examples of likelihood matrices obtained using sparse classification. The left figure shows likelihoods obtained with the histogram-based labeling proposed in [1]. The right figure corresponds to the proposed method. Darker colors indicate a higher likelihood, and digits can be recognized by their continuous paths along the diagonal. The horizontal axis represent time. The left right vertical axis shows the digit labels, with 'Z' representing 'zero' and the label 'sil' representing silence. The left vertical axis shows the underlying states indices. The utterance spoken is '16688'

matrix, whereas the histogram-based method produced wider paths.

3.2. Silence likelihoods

The likelihoods of silence states cannot be reliably estimated from noisy utterances using the method described above because of the following reasons. First, the signal model (1) does not require exemplars containing silence to represent segments without speech activity. Second, the weighted combination of exemplars may contain exemplars consisting of speech-to-silence transitions, so that silence states get a non-zero likelihood during speech.

In [1], we artificially boosted the silence state likelihood in each window by a value derived from the sum of the exemplar weights times an empirically derived constant. The constant proved difficult to optimize and give rise to numerous insertion and/or deletion errors. In this work, we propose an improved method for estimating the silence state likelihoods.

The silence state likelihood estimates are derived from the speech activity r_w that is measured for each window w by the sum of speech exemplar weights, $r_w = \sum_{l=1}^{J} x^s_{w,j}$. The measure is normalized between 0 and 1 over each utterance, so that 0 corresponds to silence and 1 to full speech activity. For each frame τ we use the speech activity r_τ measured in window w centered around frame t. The adjusted activity measure $\hat{r}_\tau$ is obtained by applying a shifted and scaled logistic function

$$\hat{r}_\tau = \frac{1}{1 + \exp(-\alpha r_\tau + \beta)} \tag{8}$$

to sharpen the distinction between speech and silence areas. Parameters α and β were optimized using development data (see Section 4) for each window length separately.

At each frame, the speech state likelihoods are multiplied by a common factor so that their sum equals $\hat{r}_\tau$. Similarly, silence states are scaled as a group until they add up to $1 - \hat{r}_\tau$. Consequently, the original ratio between speech and silence likelihoods is replaced by one defined by $\hat{r}$, and the likelihoods in each frame sum to unity.

3.3. Finding the most likely state transition path

After obtaining the state likelihoods for the entire utterance, we use the Viterbi algorithm to find the state sequence that maximises likelihood. With the state labels obtained from a forced alignment with a conventional HMM-based recognizer, we use the Viterbi back-end of that same recognizer to carry out the Viterbi search.

4. Experiments

For our recognition experiments we used material from testsets 'A' and 'B' of the AURORA-2 corpus [10]. Testset 'A' comprises 1 clean and 24 noisy subsets, containing four noise types (subway, car, babble, exhibition hall) at six SNR values, $20, 15, 10, 5, 0$ and -5 dB. Testset 'B' contains four different noise types (restaurant, street, airport, train station). Each subset contains 1001 utterances with one to seven digits '0-9' or 'oh'. To reduce computation times, we used a random, representative subset of 10% of the utterances (i.e. 400 utterances per SNR level). Acoustic feature vectors consisted of Mel frequency magnitude spectra, spanning $K = 23$ bands with a frame shift of 10 ms and a frame length of 25 ms.

For each window length we created a dictionary of 4000 noise and 4000 clean speech exemplars by randomly selecting exemplars from the noise and clean speech in the multicondition training set. The multicondition training set of AURORA-2 contains 8440 utterances with the same noises as in testset 'A', at SNR values SNR $= 20, 15, 10, 5$ dB. The spectrograms were reshaped to vectors and subsequently added as the columns of the dictionary $\mathbf{A}$ as described in Section 2. The dictionary $\mathbf{A}$ was normalized by fixing the Euclidean norm to unity along both dimensions. Finally, each observation $\mathbf{y}$ was scaled using the normalization matrices applied to $\mathbf{A}$.

HMM-state based labels of the exemplars were obtained via a forced alignment with the orthographic transcription using the HMM-based recognizer described in [11]. Viterbi decoding was done using the back-end of this, and the same (noise robust) decoder was also used for our baseline recognition experiments. Digits were described by 16 states with an additional 3-state silence word, resulting in a $Q = 179$ dimensional state-space.

We carried out recognition experiments at three window lengths: $T \in \{10, 20, 30\}$ frames. Recognition accuracies were averaged over the four noise types at each SNR level. The speech decoding system was implemented in MATLAB. To obtain results with histogram-based labeling we used the same configuration as in [1].

Silence likelihood parameters were optimized for each window length by maximising recognition accuracy on a develop-

Table 1: Word recognition accuracy at several window lengths and SNR's. The first row displays the baseline accuracy as obtained by a noise robust recognizer [9] that is based on missing data technique. The rows denoted by 'HB' denote results obtained using histogram-based labeling, the rows denoted by 'TD' represent results obtained using the proposed time-dependent labeling.

SNR [dB]		clean	20	15	10	5	0	-5
baseline		99.7	97.9	95.5	91.4	82.6	62.1	17.1
HB	T=10	95.5	93.8	92.7	90.2	83.8	69.5	41.0
	T=20	93.5	92.3	91.9	88.8	83.8	72.0	49.3
	T=30	89.5	88.4	88.0	85.5	82.6	74.9	55.8
TD	T=10	96.2	95.3	94.4	92.1	84.7	71.2	39.6
	T=20	96.6	95.8	94.8	92.7	88.8	78.1	53.1
	T=30	94.7	93.4	93.3	92.2	89.9	79.5	56.7

(a) Test set 'A'

SNR [dB]		clean	20	15	10	5	0	-5
baseline		99.7	95.3	91.2	84.3	70.4	40.2	12.2
HB	T=10	95.5	93.7	90.4	84.6	73.5	50.6	21.2
	T=20	93.5	91.6	88.6	80.8	69.1	45.1	23.3
	T=30	89.5	87.2	85.2	80.4	71.8	54.8	32.4
TD	T=10	96.2	94.7	93.6	87.9	78.4	57.1	27.4
	T=20	96.6	95.3	93.7	89.9	82.7	63.1	35.7
	T=30	94.7	93.5	93.2	90.1	85.7	67.5	37.6

(b) Test set 'B'

ment set created from unused testset 'A' and 'B' material. The development set consists of 100 utterances from each SNR and each noise type, 4800 utterances in total.

5. Results and Discussion

From the results in Table 1 we can deduce that the use of the new time-dependent labeling, in combination with the new silence likelihood calculation, leads to substantial improvements in recognition accuracy. With the exception of a single condition, SNR -5 dB on testset 'A' with T=10, the results improve for across conditions by up to 18% absolute at SNR 0 dB, T=20 on testset 'B'.

The results at T=20 and T=30 have benefited the most from the use of time-dependent labeling. The reason for this is that with increasing the length of the exemplar, it becomes increasingly difficult to accurately model the state information in exemplar using a single histogram. Since the time-dependent labelling models the state changes over time within the exemplar, the length of the exemplar no longer poses a disadvantage for decoding. Although the window length T=30 still achieves the highest accuracies at SNR 10 dB and lower, T=20 now performs even better than T=10 at higher SNRs.

The decrease in accuracy at high SNR's, compared to the baseline noise robust recognizer [9] based on missing data technique, is due to a number of reasons, such as the limited number of exemplars in the dictionary. Preliminary research shows that this drop in performance can be relatively easily be avoided by several approaches, one of which the combination of likelihoods with those obtained with a conventional recognizer. As it is, the method has great potential since we can observe that the proposed method works much better for SNR's below 10 or 15 dB., depending on the testset, with differences of up to $\approx 40\%$ at SNR -5 dB. Moreover, the proposed time-dependent labeling allows extending the basic model (1) to convolutive speech bases [12], which results in more compact, shift-invariant dictionaries.

6. Conclusions

We proposed a state-based labeling for speech exemplars spanning multiple frames to model the time-varying speech information within exemplars. Using these labeled exemplars, we did noise-robust exemplar-based ASR by decomposing noisy speech as a linear combination of speech and noise exemplars. The weights of the linear combination of exemplars were then used together with the state-labeling to provide noise-robust state likelihoods. Experiments on AURORA-2 revealed the state-based labeling works up to 18% better than a histogram-based labeling method previously used.

7. Acknowledgements

Tuomas Virtanen and Antti Hurmalainen have been funded by the Academy of Finland. The research of Jort F. Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program.

8. References

[1] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[2] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[3] G. S. V. S. Sivaram, S. K. Nemala, M. Elhilali, T. D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[4] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[5] M. Van Segboeck and H. Van hamme, "Unsupervised learning of time-frequency patches as a noise-robust representation of speech," *Speech Communication*, vol. 51, no. 11, 2009.

[6] B. Schuller, F. Weninger, M. Wllmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[7] J. Gemmeke, H. Van hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 272–287, 2010.

[8] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.

[9] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. ICASSP*, Montreal, Quebec, Canada, May 17–21 2004, pp. 213–216.

[10] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ASR2000 Workshop, Paris, France*, 2000, pp. 181–188.

[11] M. Van Segbroeck and H. Van hamme, "Robust speech recognition using missing data techniqies in the prospect domain and fuzzy masks," in *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing*, Las Vegas, USA, 2008.

[12] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, 2007.