Exemplar-based Recognition of Speech in Highly Variable Noise

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

[1]Department of Signal Processing, Tampere University of Technology, Finland
[2]Department ESAT, Katholieke Universiteit Leuven, Belgium

antti.hurmalainen@tut.fi, katariina.mahkonen@tut.fi,
jgemmeke@amadana.nl, tuomas.virtanen@tut.fi

Abstract

Robustness against varying background noise is a crucial requirement for the use of automatic speech recognition in everyday situations. In previous work, we proposed an exemplar-based recognition system for tackling the issue at low SNRs. In this work, we compare several exemplar-based factorisation and decoding algorithms in pursuit of higher noise robustness. The algorithms are evaluated using the PASCAL CHiME challenge corpus, which contains multiple speakers and authentic living room noise at six SNRs ranging from 9 to -6 dB. The results show that the proposed exemplar-based techniques offer a substantial improvement in the noise robustness of speech recognition.

Index Terms: automatic speech recognition, exemplar-based, noise robustness, sparse representation

1. Introduction

While Automatic Speech Recognition (ASR) has been under intensive research for decades, its widespread adoption is still being delayed by practical issues. One of the primary problems is varying background noise. Conventional ASR systems, based on frame level Gaussian Mixture Models (GMMs), suffer significant quality degradation when spectral features become corrupted by noise. Joint modelling of the target speech and noise in the recognizer, [1], feature compensation [2], and missing data techniques [3] have been suggested to overcome this problem. Meanwhile, there are alternative routes, which no longer employ GMMs to discover the underlying speech content.

In previous work [4, 5], we have described an *exemplar-based* recognition framework, where noisy speech is represented as a combination of multi-frame speech and noise spectrogram segments, *exemplars*. The framework can be used for signal or feature enhancement, but the best results have been achieved by using exemplar labels, which directly reveal the phonetic content of an utterance via their activation weights. In this paper, we explore the effectiveness of the exemplar-based framework on highly corrupted speech using the PASCAL CHiME challenge data, in which the speech is not only reverberated, but also contains phonetically close keywords and highly variable background noise events.

Concerning our framework, the CHiME data provides a few interesting options, which were not present in the previous experiments carried out on the AURORA-2 database. First, the data is stereophonic and high quality. Second, the utterances to be recognised can be observed within their neighbouring noise context. Finally, the identity of the speaker is known at the moment of recognition, so speaker-dependent speech exemplars can be reliably employed.

The rest of the paper is organised as follows. The general concepts of our exemplar-based approach are described in Section 2. The experimental setup, including the CHiME database, feature extraction and parameter settings of the baseline system are presented in Section 3. The baseline exemplar-based recognition results are shown and discussed in Section 4. Experiments with two variants; the use of matrix deconvolution (NMD) and the use of regression to learn the mapping between words and exemplars, are described in Sections 5 and 6, respectively. The overall discussion of our findings is presented in Section 7, and the summary and conclusions in Section 8.

2. Recognition with speech and noise exemplars

Sparse representations have received increasing attention in several applications, including image and audio signal processing. The key concept is that many natural signals can be described as a linear combination of only a few atoms. Enforcing sparsity prevents overfitting with too many elements. By allowing only a small number of activations, we can expect to find the few dictionary atoms, which best explain the mixed signal.

In noise robust speech recognition, it has been proposed that speech may be described as a sparse linear combination of *exemplars*, and that noisy speech can likewise be described as a combination of noise and speech exemplars [5, 6, 7]. When a noisy utterance is represented using these components, the activations of speech exemplars, together with knowledge of the words they represent, can be used to recognise the underlying utterance.

2.1. Sparse representation of noisy speech

The base element of our sparse representation is an *exemplar*, a $B \times T$ spectrogram block of B spectral magnitudes of speech or noise in T consecutive frames, extracted from training data. The exemplars are indexed by variable e. To simplify the notation, the columns of each spectrogram matrix are stacked into vector $\mathbf{a}_e$ of length $B \cdot T$. The E exemplars are gathered into the columns of matrix $\mathbf{A}$ to form a *basis* or *dictionary*.

The utterance to be recognised is similarly converted to spectral features. A length T observation window is concatenated into vector $\mathbf{y}$. The observation window is represented as a linear combination of exemplars,

$$\mathbf{y} \approx \sum_{e=1}^{E} \mathbf{a}_e x_e, \tag{1}$$

where x_e is the weight or *activation* of each exemplar.

In the baseline exemplar-based recognition system we em-

ploy an algorithm referred as 'NMF' (*Non-negative Matrix Factorisation*) to find the non-negative and sparse activations. The vector $\mathbf{x}$ of all activations x_e in Equation 1 can be determined simultanously for multiple observation vectors stored in columns of matrix $\mathbf{Y}$, each producing its own column to the total activation matrix $\mathbf{X}$. The matrix equation to be solved thus becomes $\mathbf{Y} \approx \mathbf{AX}$.

We obtain the non-negative activation matrix $\mathbf{X}$ while minimising the Kullback-Leibler divergence and introducing an sparsity-inducing L_1 penalty for non-zero activations by using the update rule

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{Y}/(\mathbf{AX}))}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + \mathbf{\Lambda}}. \quad (2)$$

Here $\otimes$ denotes elementwise multiplication. Matrix divisions are also elementwise. $\mathbf{1}$ is an utterance-sized all-ones matrix. $\mathbf{\Lambda}$ is the sparsity penalty matrix, defined for each activation entry.

For recognition of utterances of arbitrary length T_{utt}, we process the utterance in $W = T_{\text{utt}} - T + 1$ overlapping feature windows with a step of one frame between windows. Because the middle frames are estimated several times in consecutive windows, averaging is applied to the likelihoods of the next step to compensate for this. For a thorough description of this factorisation method, see [4]. An alternative method for handling temporal continuity, referred as *Non-negative Matrix Deconvolution* (NMD), is presented in Section 5.

2.2. Recognition

To decode the signal, we create a $Q \times T_{\text{utt}}$ *likelihood matrix* $\mathbf{L}$, where each entry $\mathbf{L}_{q\tau}$ denotes the probability of speech state q $(1 \ldots Q)$ in frame τ $(1 \ldots T_{\text{utt}})$. This is generated using *conversion matrices* $\mathbf{B}_t$ $(Q \times E)$, which describe the linear mapping of exemplars to states for each frame t of the exemplars. In our baseline system, we use binary labelling of dictionary exemplars for the conversion. In each exemplar frame only one state is labelled to be active. The matrices need not to be binary, though. in Section 6 we will experiment with a technique to *learn* the conversion matrices in order to take into account dependencies between exemplar activations.

After generating the whole matrix $\mathbf{L}$ as described in [4], each of its columns (representing state likelihoods in one frame) is normalised to unitary sum. The matrix is then decoded using a Viterbi algorithm and trained transition parameters.

3. Experimental setup

3.1. The CHiME database

The PASCAL 'CHiME' Speech Separation and Recognition Challenge [8] is designed to address some of the problems occurring in real world noisy speech recognition. The challenge data is based on the GRID corpus [9], where 34 speakers read simple command sentences. These sentences are of form *verb-colour-preposition-letter-digit-adverb*. There are 25 different 'letter' class words and 10 different digits. Other classes have four word options each. In the CHiME recognition task, the final score is the percentage of correctly recognised 'letter' and 'digit' keywords.

CHiME utterances simulate a scenario, where sentences are spoken in a noisy living room. The original, clean speech utterances are reverberated according to the actual room response, and then mixed to selected noise sections, which produce the desired SNR mixture level for each noisy set. The noisy sets have target SNR levels of 9, 6, 3, 0, -3 and -6 dB.

For modelling/training, there are 500 reverberated utterances per speaker (no noise), and six hours of background noise data. The development and test sets consist of 600 mixed-speaker utterances at each SNR level, Additionally, noiseless (only reverberated) development utterances are available. Development and test utterances are both given in a strictly end-pointed format, but also as embedded signals within their noise context. All data is stereophonic and has a sampling rate of 16 kHz.

3.2. Feature extraction

For the features of our framework, we used spectral magnitudes of Mel bands. These were calculated from partially overlapping 25 ms frames with a shift of 10 ms between frames. 26 bands were used for the 16 kHz signal (Nyquist frequency 8 kHz), which matches the number of bands used for the default CHiME MFCC models. Features were extracted separately for both stereo channels and concatenated, thus effectively doubling the number of feature bands.

3.3. Speech exemplars

We used 5000 speech and 5000 noise exemplars for each window length T, adding up to $E = 10000$ total entries. We created two different types of speech dictionaries: a speaker-dependent and a speaker-independent one. First, an initial speech dictionary was created for each speaker, based on a 60% subset of the noiseless speech training utterances, by extracting exemplars with a random frame shift of 4 to 8 frames. This produced approximately 10000–17000 partially overlapping exemplars per speaker and window length. For the speaker-dependent dictionaries, each initial dictionary was reduced to a fixed size of 5000 exemplars by selecting exemplars such that there is a maximally flat coverage between words. (In the original dictionaries, words from classes with fewer options are over-represented due to more frequent appearance in the training set.)

A speaker-independent dictionary was created for each window length, this time by selecting 147–148 (5000/34) exemplars from each full speaker-dependent dictionary with similar word probability flattening. These were then combined to a single 5000 exemplar dictionary per window length.

In addition to storing the spectral feature data, state labels were assigned to the speech exemplars by using transcriptions acquired by forced alignment. Alternatively, the state information was learnt by factorising the remaining 40% of training files and finding the mapping as described in Section 6.

3.4. Noise exemplars

The selection of noise exemplars has a central role in the separation quality of factorisation algorithms. If no matching noise is found, separation results become unpredictable. Initially, we created two different types of noise dictionaries. In the first, 5000 noise exemplars were randomly extracted from the provided background noise data. In the second, 5000 noise exemplars were selected by sampling the neighbourhood of embedded utterances to both directions with a shift of 4 to 7 frames, excluding locations where other test utterances were embedded.

Experiments using the development set (not shown) indicated that using the adaptive noise dictionary yields a 1–4% improvement in recognition accuracy compared to the fixed noise dictionary. In this paper, we will only report results obtained using adaptive noise.

Table 1: Results of the baseline exemplar-based recogniser on the test set. The rows refer to different exemplar sizes. CHiME GMM baseline results are also shown. The best result at each SNR level is highlighted.

SNR (dB)	9	6	3	0	-3	-6
CHiME baseline	**82.1**	70.8	61.3	52.0	39.8	34.7
$T = 10$	69.9	66.0	58.7	52.4	42.9	37.8
$T = 20$	77.3	72.8	**68.2**	**62.7**	51.1	44.0
$T = 30$	76.0	**73.5**	**68.2**	61.8	**52.7**	**44.7**

(a) Speaker-independent results

SNR (dB)	9	6	3	0	-3	-6
CHiME baseline	82.4	75.0	62.9	49.5	35.4	30.3
$T = 10$	91.3	88.3	85.8	80.8	71.4	62.3
$T = 20$	**91.6**	**89.2**	**87.6**	**84.2**	74.7	68.0
$T = 30$	88.8	88.1	86.3	82.9	**75.1**	**68.3**

(b) Speaker-dependent results

3.5. Processing test utterances

For factorisation, each utterance was read from the endpointed ('isolated') file, and converted into Mel features. After choosing the appropriate speech and noise basis for the utterance, they were reweighted together to equal Euclidean norm over Mel bands and exemplars. Band weights from the combined dictionary were then applied to the utterance features.

The NMF penalty matrix $\mathbf{\Lambda}$ used in finding a sparse representation can be set for each activation entry separately. We used two different values, one for speech exemplars and another for noise. The values were tuned by factorising a subset of development utterances with partially adaptive, speaker-dependent bases and exemplar size $T = 20$. The penalty values were set as 2.0 and 1.7 for speech and noise exemplars, respectively. Generally speaking, higher values of $\mathbf{\Lambda}$ produce better recognition rates at high SNRs, while lower ones lead to better performance at low SNRs. We selected values, which give a slight emphasis to the noisy end. The same sparsity values were used throughout all experiments.

For state representation, we used the same model as in the CHiME baseline recogniser. Each word is modelled with 4–10 successive states, and the whole system uses in total 250 states. The activations were mapped to state likelihoods as explained in section 2.2. Utterances were decoded using the HVite binary of the HTK toolkit, modified to pick its state likelihoods directly from the generated matrix $\mathbf{L}$ instead of evaluating state GMMs.

4. Baseline system results

The results of the baseline exemplar-based recogniser are presented in Table 1. Three different window lengths, $T = 10$, 20 and 30 are shown, as well as results for both speaker-dependent and speaker-independent systems. The GMM-based CHiME baseline recognition results are also shown. When comparing the results, note that the baseline system uses mono features without noise compensation other than cepstral mean normalisation.

In general, it is clear that the exemplar-based recognition system outperforms the baseline GMM system in almost all conditions, especially when using speaker-dependent speech dictionaries. The lower performance of speaker-independent dictionaries ensues because a mixed speech dictionary only has a very limited number of exemplars to match a certain speaker, while at the same time it has a larger chance of matching to speech features in the background noise, produced by people in the living room or by various entertainment appliances. Interestingly, the speaker-independent GMM-based system was more noise robust at low SNRs, possibly because the trained Gaussians have a larger variance and thus match corrupted speech features better.

Like in experiments on AURORA-2 [4, 5], using an exemplar size of $T = 10$ was found suboptimal at low SNRs, because not enough time context can be exploited. $T = 20$ generally turned out equal or superior to $T = 10$. Exemplar size $T = 30$ is the most robust against noise, but performs worse at high SNRs. As the exemplar size increases, the dimensionality of feature vectors grows, and it becomes more difficult to find a matching linear combination of speech exemplars. Using a higher number of exemplars may alleviate this effect, at the cost of increased computational complexity.

5. Non-negative matrix deconvolution

As a first variant of the baseline exemplar-based recognition system, we use *Non-negative Matrix Deconvolution* (NMD) rather than NMF to obtain sparse representations of noisy speech. NMD is a name given to an alternative method to handle temporal continuity between frames. The algorithm has also been called *convolutive sparse coding* [10].

While not a deconvolution algorithm in the traditional sense, the name reflects the principle that a reconstructed utterance is represented as a convolution between activations and exemplars. This means that all the activations jointly form the estimated utterance matrix. A few activations at specific temporal locations are typically enough to represent the utterance features. There are no independent estimates or averaging like in the sliding window NMF. For the convolutive update algorithm and comparison of behaviour, see [11].

The results for NMF and NMD algorithms are shown in Table 2. Both methods employ adaptive noise dictionaries, speaker-dependent speech dictionaries and 300 iterative updates. In NMF, the speech exemplar activations were normalised to unitary sum in each window. In NMD, no normalisation was performed. These choices have been found recommendable in earlier work [4, 11].

In these results, NMF produces slightly yet significantly better recognition rates in all conditions. This is surprising, because on AURORA-2 we observed the opposite: NMD outperformed NMF. Especially the degradation of NMD at $T = 30$ is unexpected, because on AURORA-2 it was the best performing exemplar size [11].

One possible reason is that factorisation parameters were optimised using NMF. Because the full optimisation process is computationally heavy, the same parameters were applied directly to NMD. Therefore the results may favour NMF. We can also speculate, that the closely related keywords in CHiME are prone to occasional misclassifications in sparse activations. As there is more averaging over independent estimates in NMF, the chance of errors in the final estimate is smaller than in NMD. Because a 1–2% drop was already present in the cleanest end of both keyword classes, we can suspect a problem with word recognition itself, not the noise robustness of NMD.

Table 2: Comparison of NMF and NMD factorisation algorithms in speaker-dependent recognition. The rows refer to different exemplar sizes. The best result at each SNR level is highlighted.

SNR (dB)	9	6	3	0	-3	-6
CHiME baseline	82.4	75.0	62.9	49.5	35.4	30.3
$T = 10$	91.3	88.3	85.8	80.8	71.4	62.3
$T = 20$	**91.6**	**89.2**	**87.6**	**84.2**	74.7	68.0
$T = 30$	88.8	88.1	86.3	82.9	**75.1**	**68.3**

(a) NMF

SNR (dB)	9	6	3	0	-3	-6
CHiME baseline	82.4	75.0	62.9	49.5	35.4	30.3
$T = 10$	88.3	85.9	83.3	78.8	69.1	59.8
$T = 20$	**90.5**	**88.6**	**87.0**	**81.3**	**72.1**	**65.9**
$T = 30$	87.2	86.1	84.0	79.9	70.6	63.3

(b) NMD

6. Mapping from activations to likelihoods

In our baseline system, the mapping from activations to word state likelihoods is based on labels of dictionary items, which have been obtained by forced alignment. However, in label-based mapping of word models there is the inherent problem that phonetically similar features may appear in different contexts. A factorisation algorithm (NMF or NMD) selects the exemplars with best fitting spectral features, while their labels may occasionally suggest a misleading word identity. Such an error will easily result in a misclassification.

We tested an alternative approach, where the mapping was not assigned according to dictionary labels, but *learnt* using regression algorithms on factorised training data labelled by forced alignment. Labels were assigned to a 40% subset of the training set for this purpose. Then a regression algorithm was used to discover optimal mapping matrices between activation vectors and target states.

We used two different regression algorithms, *Ordinary Least Squares* (OLS) and *Partial Least Squares* (PLS) to learn the mapping from activations to likelihoods. OLS is straightforward minimisation of the L_2 error term in mapping. PLS (also known as Projection to Latent Structures) uses an internal, usually lower dimensioned space. The original coordinates are rotated in input and output to the internal space, where the true mapping is optimised. PLS can tolerate a collinearity of input data, contrary to OLS. For details, see [12].

The outcome of the recognition with different likelihood generation methods is shown in Table 3. Results are listed for recognition with binary labels, and OLS/PLS-trained mapping. Speaker-independent results are included, because they provide interesting insight to scenarios where flaws of the original system can be countered with learning.

In speaker-independent recognition, uniform improvements of 4.3–14.1% (absolute) can be seen over the use of binary labels. In these dictionaries, very few instances of each word are present for a specific speaker. This seems to result in numerous misclassifications due to exemplars from other words being activated instead. When the conversion matrices are learnt — in this case from a large amount of training material — the actual correspondence of each exemplar can be discovered with convincing results. Possibly for the abundance of training material coming from all speakers, OLS is mostly superior to PLS.

The speaker-dependent results are more mixed. Here the dictionaries only cover one speaker at a time, and thus can include a broad representation of all words and states. In fact, the reduction algorithm did not remove any of the letter and digit exemplars gathered from the training material, because they all fit in the 5000 exemplar dictionaries. It is also worth noting, that in this scenario the regression matrices were only trained from the speaker's own training subset (200 utterances), which is quite limited regarding keyword appearance. Under this limited training data, the performance of all methods was mostly similar, unlike in the speaker-independent case.

7. Discussion

The CHiME challenge database provided some new insight to the applicability of our exemplar-based methods. Overall, the results appear very plausible. Using properly selected algorithms and parameters, our framework reduced the recognition error rates to less than half of the CHiME baseline system at all SNRs, in many cases even by significantly larger a margin. We also achieved improvements in noise robustness over our previous work on AURORA-2. These gains can be partially attributed to the characteristics of CHiME, which allow construction of accurate dictionaries for both speech and noise.

When the speaker identity is known and thus matching speech exemplars can be selected, correct phonetic features can be picked out reliably even in the presence of other voices. Our speaker-dependent results were significantly better than the speaker-independent ones. Using GMMs the difference was not so clear. Regarding noise dictionaries, we found out that adaptive noise exemplar selection can yield high separation quality under varying noise conditions. Previously there were some concerns over the feasibility of generating a generic noise dictionary using a practically manageable number of exemplars. Our CHiME experiments confirm, that adaptive selection can be used instead of a fixed dictionary. Its implementation should be feasible in practical applications as well.

One surprising and slightly disappointing aspect was the subpar performance of NMD in comparison to sliding window based NMF. It is not certain yet, whether this is a real algorithmic difference or merely a result of insufficient parameter training in NMD. Further experiments and optimisations should be carried out to find out the true capabilities of each factorisation algorithm.

More favourable results were achieved in learnt likelihood mapping. The gains over explicitly assigned labels are positive by themselves. However, in a larger context this means that well performing likelihood mappings can be learnt even for features, which are not directly derived from any specific speech sections. In other words, we can experiment with any kind of dictionary generation methods and then find out the phonetic labels even if none were originally present.

While the separation and likelihood generation algorithms of our framework have already been improved, more attention should be paid to optimising the features and state models for maximal linguistic accuracy. The CHiME data illustrates, how some closely related words can be difficult to tell apart even under favourable conditions. Although noise robustness is a crucial aspect in practical ASR systems and our framework has

Table 3: Comparison of the recognition with three different likelihood generation methods on the test set. In addition to binary labels, OLS and PLS regression are evaluated. The best result at each SNR level and for each exemplar size is highlighted.

SNR	(dB)	9	6	3	0	-3	-6
CHiME	baseline	82.1	70.8	61.3	52.0	39.8	34.7
$T = 10$	labels	69.9	66.0	58.7	52.4	42.9	37.8
	OLS	**84.3**	**77.8**	**71.4**	**65.3**	56.4	48.6
	PLS	82.1	77.1	71.0	64.0	**57.0**	**49.3**
$T = 20$	labels	77.3	72.8	68.2	62.7	51.1	44.0
	OLS	**85.2**	**80.5**	**78.7**	**71.1**	**60.2**	**51.5**
	PLS	82.9	78.8	74.8	70.1	59.5	50.6
$T = 30$	labels	76.0	73.5	68.2	61.8	52.7	44.7
	OLS	**82.8**	**80.5**	**76.3**	**70.7**	**62.1**	**54.4**
	PLS	81.1	77.8	74.3	68.8	61.1	52.4

(a) Speaker-independent recognition

SNR	(dB)	9	6	3	0	-3	-6
CHiME	baseline	82.4	75.0	62.9	49.5	35.4	30.3
$T = 10$	labels	**91.3**	**88.3**	**85.8**	**80.8**	**71.4**	62.3
	OLS	89.8	86.8	85.0	79.7	70.1	62.7
	PLS	90.5	87.8	84.5	80.2	71.3	**63.7**
$T = 20$	labels	91.6	89.2	87.6	84.2	74.7	68.0
	OLS	91.1	**90.0**	**88.5**	**85.2**	77.6	69.2
	PLS	**91.9**	89.3	88.2	85.0	**78.6**	**69.6**
$T = 30$	labels	88.8	**88.1**	86.3	82.9	75.1	68.3
	OLS	88.8	86.0	**86.4**	**83.3**	76.1	**69.2**
	PLS	**89.1**	85.7	84.8	82.4	**77.2**	68.8

(b) Speaker-dependent recognition

shown significant advances in achieving it, the ultimate goal of maximally accurate recognition of speech itself should not be forgotten or compromised. Proper phonetic state models should be introduced instead of the current word models to avoid multiple meanings between similar features, and to make large vocabulary recognition feasible.

8. Conclusions

Exemplar-based methods were presented for recognition of speech in highly variable real world noise. The main framework and its variants were evaluated using the CHiME challenge database, which covers actual living room noise at multiple SNRs. We achieved recognition rates of over 91% at 9 dB, and close to 70% at -6 dB. Long temporal context with 200 ms exemplars, speaker-dependent speech dictionaries and adaptive noise dictionary gathering were found the best options for recognition of noisy speech.

Two separation algorithms, non-negative matrix factorisation and -deconvolution were used for determining the exemplar activations from Mel-scale spectral magnitude features. In these experiments, factorisation of overlapping windows independently from each other performed better than deconvolutive separation of whole utterances at once.

Learning the mappings from exemplar activations to state likelihoods using OLS and PLS regression was proposed. These algorithms were compared to strict binary labels acquired from forced alignment. Highest gains were seen in speaker-independent recognition. The original binary labels produced unreliable results, while mappings learnt from large training data improved the recognition rates by 4–14% (absolute). In speaker-dependent recognition the differences were small.

The results surpassed significantly both the CHiME baseline results and our previous AURORA-2 recognition rates. While the noise robustness of our system is already relatively high, parameter optimisation and better speech models would help in improving the overall recognition quality even further.

9. Acknowledgements

Antti Hurmalainen has been funded by the Academy of Finland. The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049.

10. References

[1] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition: Graphical modeling approaches," *IEEE Signal Processing Magazine*, vol. 27, no. 6.

[2] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Atlanta, USA, 1996.

[3] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, September 2005.

[4] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Accepted for publication in IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[5] T. Virtanen, J. F. Gemmeke, and A. Hurmalainen, "State-based labelling for a sparse representation of speech and its application to robust speech recognition," in *Proceedings of INTERSPEECH*, Makuhari, Japan, 2010.

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

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

[8] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proceedings of INTERSPEECH*, Makuhari, Japan, 2010.

[9] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120(5), 2006.

[10] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, München, Germany, 2004.

[11] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Prague, Czech Republic, 2011.

[12] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica Chimica Acta*, vol. 185, no. 1, 1986.