EUSIPCO 2013 1569744587

ACQUIRING VARIABLE LENGTH SPEECH BASES FOR FACTORISATION-BASED NOISE ROBUST SPEECH RECOGNITION

Antti Hurmalainen, Tuomas Virtanen

Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland

ABSTRACT

Studies from multiple disciplines show that spectro-temporal units of natural languages and human speech perception are longer than short-time frames commonly employed in automatic speech recognition. Extended temporal context is also beneficial for separation of concurrent sound sources such as speech and noise. However, the length of patterns in speech varies greatly, making it difficult to model with fixed-length units. We propose methods for acquiring variable length speech atom bases for accurate yet compact representation of speech with a large temporal context. Bases are generated from spectral features, from assigned state labels, and as a combination of both. Results for factorisation-based speech recognition in noisy conditions show equal or better separation and recognition quality in comparison to fixed length units, while model sizes are reduced by up to 40%.

Index Terms— Spectral factorization, speech recognition, noise robustness

1. INTRODUCTION

Speech contains phonetic units of varying lengths, ranging from single phones to their combinations, syllables, words and complete phrases. Statistical analysis of speech reveals correlation in its temporal behaviour spanning hundreds of milliseconds, decreasing gradually with no strict upper limit [1]. Meanwhile, physiological studies and listening tests have shown that temporal modulations at under 12 Hz (period of 83 ms or more) are crucial for speech intelligibility [2].

Conventional automatic speech recognition (ASR) systems typically use frames of approximately 25 ms as their features, and Markovian state transition models which only consider temporal context of one frame. The approach is computationally efficient and sufficient for single phone classification, but fails to model the long term temporal behaviour motivated by natural speech structures and human hearing. Especially in noisy conditions short-term spectra become unreliable as features for classification. Separating and recognising sources from a single frame is often an ill-posed problem. While partial alleviation can be achieved by including delta and acceleration features to frame spectra, the context still remains limited, and extended temporal connectivity actually violates the Markovian model assumption [1]. Due to these limitations and the need for more robust models, there is increasing interest towards long context spectrogram modelling in ASR [3].

Several approaches have been proposed for increasing the context of speech models. TRAPs features observe long term temporal behaviour of a few spectral bands [4]. HAC models quantise frame level audio events into classes and form histogram vectors summarising the events in variable length words [5]. Phonetic segmentation of speech has been discussed and demonstrated in literature [6], although evaluation has usually consisted of comparison to manual segmentation with no application to ASR. Longest segment matching has been applied to dereverberation [7] and robust ASR [8]. Increased context has also been used in deep belief networks with optimal results gained at contexts of 110–270 ms [9].

Using spectro-temporal atoms spanning 200–300 ms has been shown to provide high separation quality and noise robustness with methods based on *non-negative matrix factorisation* (NMF) [10, 11, 12]. However, the exact choice of window length has proven difficult. Increasing the context will improve robustness. On the other hand, it increases the complexity of modelled spectro-temporal patterns, thus requiring more atoms for the same data. Furthermore, fixed atom length does not correspond to the large variation of acoustic units occurring in real world speech and noise.

While virtually all studies on NMF thus far have concentrated on fixed atom length models, more recently variable length modelling has also been proposed. Yılmaz et al. used combination of factorisation passes with multiple fixed length dictionaries [13]. Although a promising step towards variable length modelling, using multiple large dictionaries may prove impractical. Meanwhile, Wang and Tejedor have proposed a model for employing different atom lengths simultaneously in convolutive NMF (also known as NMD) [14], and presented an introductory experiment on two-speaker separation.

In this work we extend variable length NMD modelling to robust ASR, and propose methods for acquiring compact speech bases with a preference for long context, yet able to model units of any length. We employ two data sources for finding units; unannotated spectral features, and state labels acquired from a language model via forced alignment. The two sources are also used in conjugation. Models are evaluated in a noisy ASR task using the 1st CHiME Challenge corpus [15]. Factorisation-based representation is used for

T. Virtanen has been funded by the Academy of Finland, grant #258708.

feature enhancement for an external back-end, and for ASR directly from atom activations. First we introduce the fundamentals of spectral factorisation. The proposed acquisition method is presented in Section 3. In Section 4 we describe the evaluation set-up and experiments. Results are listed and discussed in Section 5, whereafter we conclude in Section 6.

2. SPECTROGRAM FACTORISATION

In spectrogram factorisation, the goal is to model a mixed *observation spectrogram* $\mathbf{Y}$ as a sum of separated source spectrogram estimates, which in robust ASR comprise speech $\mathbf{\Psi}^{\mathrm{s}}$ and noise $\mathbf{\Psi}^{\mathrm{n}}$. The dimensions of each utterance spectrogram are $B \times T_{\mathrm{utt}}$, where B is the number of spectral bands and T_{utt} is the number of frames. The estimates are constructed by weighted summing of *atom spectrograms* $\mathbf{A}_l$ ($B \times T_l$). Atoms are indexed by l from 1 to *basis size* L. Whereas in earlier work the *atom length* T_l has been a constant [11, 12], in this work we allow it to vary between atoms.

Each utterance spectrogram estimate $\mathbf{\Psi}$ is a convolutive sum of atom spectrograms, weighted by a $L_G \times W$ *activation matrix* $\mathbf{X}$. L_G is the number of atoms belonging to set G of the source(s) being modelled. W is the number of permitted *window indices*, equal or less than T_{utt}. The convolutive reconstruction formula for a spectrogram $\mathbf{\Psi}_G$ is

$$\mathbf{\Psi}_G = \sum_{l \in G} \sum_{t=1}^{T_l} \mathbf{A}_{l,t} \overrightarrow{\mathbf{X}_l}^{(t-1)}, \tag{1}$$

where $\mathbf{A}_{l,t}$ denotes the t^{th} frame column of atom l, $\mathbf{X}_l$ is the l^{th} row vector of $\mathbf{X}$, and operator $\rightarrow$ shifts it right by $t-1$ columns in a length T_{utt} zero-padded array to make all partial matrices to be summed $B \times T_{\mathrm{utt}}$. The method is otherwise similar to commonly used convolutive modelling [16], except that the atom length T_l can be given separately for each atom.

Assuming a pre-generated supervised basis $\mathbf{A}$, the factorisation task consists of finding the activation matrix $\mathbf{X}$ for a chosen quality function. After solving $\mathbf{X}$, it is used either for estimating source spectrograms as above, or directly as a classifier by observing the activated atoms and their supplementary label information. Both methods have been used for robust ASR. Their details can be found in earlier work [10, 12], and are also given briefly in the following sections.

3. COMPACT VARIABLE LENGTH BASES

In previous work, fixed length atoms have been acquired by sampling randomly a large amount of *exemplars* [10], or by constructing templates for each word of a small vocabulary [11, 12]. However, both methods may prove problematic when real world speech must be modelled. In order to cover spectro-temporal patterns of speech with a compact set of atoms, we propose an algorithm which aims at discovering recurring events of variable length with a preference for long units. The algorithm is based on searching for *clusters* of speech segments which match each other.

First, let us define a similarity measure c between two frame feature vectors $\mathbf{f}^{(i)}$, $\mathbf{f}^{(j)}$. The frames are considered *matching* if their c value exceeds a given threshold θ. Similarly, two *sequences* of length N, $[\mathbf{f}_1^{(i)} \dots \mathbf{f}_N^{(i)}]$ and $[\mathbf{f}_1^{(j)} \dots \mathbf{f}_N^{(j)}]$ are considered matching if all their mutual vector pairs $\mathbf{f}_n^{(i)}$, $\mathbf{f}_n^{(j)}$ match. Because the atoms in NMD are rigid with no time warping, it is crucial that sequences match throughout their duration.

We consider two different data sources for finding matches. First, we observe the spectral features of frames, denoted by $\mathbf{s}$. Spectral matching can be defined by any similarity measure, but in this work we use straightforward dot product

$$c_{\mathrm{s}}(i,j) = \mathbf{s}^{(i)} \cdot \mathbf{s}^{(j)} \tag{2}$$

between L_2-normalised spectrum vectors. The largest possible spectral similarity is thus 1.

The second method is using phonetic state labels acquired from word transcriptions with forced alignment. Each frame in training data is given a label denoting its membership in exactly one language model state q of total Q states. We define the similarity of states in frames i and j as

$$c_l(i,j) = \begin{cases} \gamma_{\mathrm{full}} & \text{if } q^{(i)} = q^{(j)} \\ \gamma_{\mathrm{part}} & \text{if } |q^{(i)} - q^{(j)}| = 1 \\ \gamma_{\mathrm{none}} & \text{otherwise} \end{cases} \tag{3}$$

The midmost 'partial match' is true if the states follow each other in a linear language model, thus allowing minor errors in alignment. Finally, the similarity measures may be combined by using a merging function. In this work the function is the sum of coefficients,

$$c_m(i,j) = c_{\mathrm{s}}(i,j) + c_l(i,j) \tag{4}$$

The relative significance of spectral and state similarity can be defined via γs and the threshold θ.

Clustering is implemented with a greedy longest-first search. Starting from the largest allowed atom length T_{max}, we find pairwise matching sequences from training data. If a sequence is found with a sufficient number of matches to other sequences, these instances form a cluster and further an atom. The contained frame ranges are flagged as taken. Then the algorithm continues clustering, reducing the atom length T when clusters of chosen size can no longer be found. Halting can be defined e.g. by the number of extracted atoms, percentage of modelled training data, or a minimum atom length T_{min}. Although the greedy algorithm does not guarantee global maximisation of atom lengths, it is practically viable and produces a basis of recurring spectral patterns in a descending order of length and frequency of occurrence.

4. EXPERIMENTAL SET-UP

4.1. Data set and features

For the experiments, we used the GRID-based 1st CHiME Challenge corpus [15]. Its speech consists of six-word com-

mand utterances following a linear *verb-colour-preposition-letter-digit-adverb* grammar. Word classes have cardinalities 4/4/4/25/10/4 respectively, totalling to 51 words. The task is to recognise 'letter' and 'digit' keywords. There are 34 speakers, and a 500-utterance training set is provided for each. Speaker identity is assumed known in recognition. Noisy development and test sets both contain 600 utterances mixed with highly non-stationary room noise at six SNRs ranging from +9 to -6 dB. All audio data contains room reverberation.

All binaural source audio was converted into 40-band mel-spectral features using 25 ms frames with 10 ms shift, and averaged into mono. Spectral bands were equalised with fixed band weights derived from training data [12]. Default CHiME language model comprising 250 sub-word states and its forced alignment were used to assign state labels to frames.

4.2. Frame correlation functions

Three correlation variants were used for clustering speech frames in basis acquisition:

1. Spectral features only ('spect')
2. State labels only ('label')
3. Combination of the two ('comb')

The spectral space employed mel magnitudes with square root compression and augmented delta features derived from a five-frame window [12]. Spectral similarity c_s was measured as the dot product of 2-normalised vectors. The features were chosen for invariance to absolute loudness, while retaining the temporal dynamics of speech. In purely spectral acqusition θ was set to 0.89 and state correlation c_l was 0.

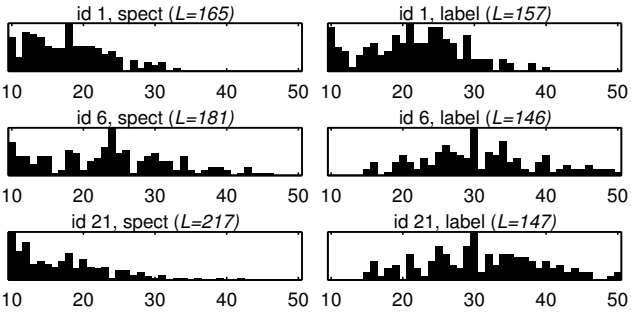
In solely label-based acquisition, c_s was in turn set to 0. Label correlation values were $\gamma_\mathrm{full} = 2$, $\gamma_\mathrm{part} = 1$ and $\gamma_\mathrm{none} = 0$. Threshold θ was set to 1 with an additional constraint that the mean correlation between sequences was over 1.8. In other words, all state pairs must correlate at least partially, and 80% of them must match perfectly. As the algorithm has no access to spectra, we require relatively strict state sequence similarity with a small allowance for fluctuations.

Combined acquisition used the same spectral correlation with θ increased to 0.92. However, c_l used $\gamma_\mathrm{full} = 0.06$, $\gamma_\mathrm{part} = 0.03$ and $\gamma_\mathrm{none} = 0$. Parameters were tuned in 0.01 steps using development data.

4.3. Basis acquisition

After defining the frame correlation functions, speech bases were acquired from training data for each speaker separately as follows. Starting from T_{max}, all pairwise matching sequences were searched from training data. Because in GRID data each word is chosen randomly from its class and no word transition is more likely than another, we restricted learning to clusters modelling a single word each. A cluster was selected if it contained at least 25% of the modelled word's instances. At each window length, all such clusters were extracted in a descending order of relative size, whereafter T was reduced by one frame. The process was halted either by reaching the minimum length T_{min}, or if at least 75% of non-silent training frames were already covered. Sequences were allowed to span over silent frames (defined by spectral energy) to model e.g. stop consonants, but not to end in one, as such cases could be modelled with a shorter atom instead.

Fig. 1. Histograms of atom lengths in selected speakers' bases for spectrum-based (left) and label-based (right) acquisition. L is the total number of atoms.



Each cluster was converted into a speech atom by averaging its mel magnitude spectrograms binwise. In addition, the preceding and succeeding 2 frames were included in atoms, because their magnitude content is implied by delta features. Original T_{max} and T_{min} were set to 46 and 6, thus final atom lengths ranged from 50 to 10 frames. A few examples of atom length histograms within individual speakers' bases are shown in Figure 1. For now, we can notice that the whole range is employed in different variants, and the distribution depends heavily on the speaker and the method. Further analysis is given later in Section 5.

A summary of the generated bases is shown in Table 1. For each generation method; spectrum-based ('spect'), label-based ('label') and combined ('comb'), we list the statistics of atom counts, total frame counts, and average atom lengths of the 34 speaker-dependent bases. Previously used fixed-length bases ('fixed') with exactly 250 length 25 atoms per speaker are included for comparison [12].

4.4. Factorisation and recognition

The factorisation and recognition framework mostly follows small basis experiments described in [12]. A joint speech+noise basis was formed from a variable number of speaker-dependent speech atoms (see Table 1), and 250 noise atoms sampled from the context of test utterances. Activation matrices $\mathbf{X}$ were solved with variable-length NMD described in Section 2 [14]. 300 iterations were used as before, and an L_1 sparsity penalty was applied for better separation and classification. Where possible, parameters were set as in the 250+250 fixed length atom experiments in [12]. Especially the fixed length noise atoms were replicated exactly to study the contribution of new speech models alone.

For decoding and recognition, we used two methods. The first is sparse classification (SC) via activation weights and

Table 1. Statistics of the 34 speaker-dependent speech bases, listed for all acquisition methods. Number of atoms in a basis, amount of contained frames, and average atom length are reported as minimum, mean and maximum values over speakers. The reference method always uses 250 length 25 atoms.

method	atom count			frame count			avg atom length		
	min	mean	max	min	mean	max	min	mean	max
spect	160	190	237	2941	3793	4659	17.0	20.0	23.6
label	135	150	181	3346	4167	5052	21.6	27.9	33.9
comb	157	182	232	3151	4027	4903	18.6	22.2	25.8
fixed	250			6250			25		

atom labels [10, 12]. In this method, a $Q \times T_{\text{utt}}$ state likelihood matrix is generated similarly to the spectrogram estimate of Equation (1) using $Q \times T_l$ label matrices assigned to speech atoms. Labels were learnt by partial training set factorisation and ordinary least squares regression between activations and utterance state content [12]. Final likelihood matrices were decoded directly using the default CHiME HMMs.

The second method is feature enhancement (FE) by using the ratio $\mathbf{\Psi}^{\text{s}}/(\mathbf{\Psi}^{\text{s}} + \mathbf{\Psi}^{\text{n}})$ of speech-only and total spectral reconstructions from Equation (1) as a time-varying filter for the original utterance spectrogram [12]. The enhanced signal was passed to a multi-condition trained robust GMM backend, previously used in [11, 12]. Details of both methods can be found in earlier work [12].

5. RESULTS AND DISCUSSION

Speech recognition and enhancement results for each modelling method are listed in Table 2 as keyword recognition rates for sparse classification (SC) and feature enhancement (FE), and signal-to-distortion ratio (SDR) of enhanced utterances measured with the BSS Eval toolkit [17]. Shown values are averages over noisy conditions and given for development and test sets separately. The first line contains baseline results for unenhanced signals. The next three lines correspond to similarity measures defined in Section 4.2 for variable length modelling. Results for previous 250+250 atom fixed-length modelling ('fixed'), and significantly larger 5000+5000 atom NMF bases ('large') are also included for comparison [12].

First, we can observe from Table 1 that in each measure of basis sizes, approximately 10–25% deviations take place between speakers from the mean to minimum and maximum values, illustrating the model's adaptivity. Mean atom count is reduced by 24.0–40.0% and mean frame count by 33.3–39.3% in comparison to fixed-length bases. Mean atom lengths vary significantly between speakers and methods. Spectral models produces more and shorter atoms than labels. Source combination generally falls inbetween.

Although the statistics ultimately depend on the similarity functions and clustering parameters, the observed trend can be justified by properties of the functions. Feature-only modelling will discover recurring spectral units, which are often shorter than whole words due to coarticulation and natural variation in pronunciation. State-only models are based on forced alignment, which always produces a similar sequence regardless of phonetic variation. It only observes variations in pacing, which are more consistent for any given speaker. This can be seen in the atom length histograms of Figure 1. Speaker 1 is fast and produces short atoms for both methods. Speaker 6 is slow and clear, hence both bases have longer atoms. Speaker 21 is relatively slow but very melodic. In this case, the feature-based atoms are shortest in the whole set, whereas state-based atoms are among the longest.

Table 2. Keyword recognition rates (%) and SDRs (dB) for unenhanced signals, proposed, and reference basis acquisition methods. Results are averages over noisy conditions from +9 to -6 dB. The best result among small basis methods (spect, label, comb, fixed) for each set is highlighted.

method	development set			test set		
	SC	FE	SDR	SC	FE	SDR
unenh	-	74.6%	-0.72 dB	-	74.7%	-0.78 dB
spect	79.4%	85.1%	7.87 dB	79.9%	85.4%	8.50 dB
label	78.3%	**85.6%**	**8.80 dB**	78.9%	**85.6%**	**8.86 dB**
comb	**79.7%**	85.3%	8.54 dB	80.3%	85.5%	8.58 dB
fixed	78.0%	84.8%	8.57 dB	**80.8%**	85.2%	8.62 dB
large	85.9%	86.7%	9.49 dB	85.8%	86.8%	9.55 dB

Regarding the separation and recognition results of Table 2, there is some variation between methods for different result metrics. While separation measured by SDR is either above or below the previous 'fixed' method, FE-based ASR results improve uniformly. Gains are small, but it should be noted that the gap to 20 times larger exemplar models ('large') is only $< 2\%$. The performance of SC is harder to analyse, because the results for development and test set differ greatly for the fixed-length model, while the proposed methods are more consistent. One contributing factor is that SC for CHiME data depends heavily on keyword modelling. In the previous model, at least four atoms per word were guaranteed, whereas the proposed method has no such constraints. In individual SNR level scores (not shown), the proposed methods had slightly lower clean end classification quality but higher robustness towards low SNRs. Separation and classification also have partially conflicting goals with the former preferring long atoms, but the latter requiring also short atoms which bear a higher risk of confusion with noise.

The main benefit of the presented method is that it can adapt to any vocabulary and speaking style, unlike the previous model which assumed long context implied by sub-word labels of small vocabulary and required defining the window length explicitly. Although a small vocabulary task was used here for simplicity of presentation and easier comparison to earlier work, we have already employed the methods — both feature- and state-based — successfully to compact modelling

of medium vocabulary speech [18]. Regarding complexity, the basis acquisition time for this task was < 30 minutes per speaker using MATLAB code and an E8400 dual-core desktop PC. For larger corpora, computation of full similarity may become slow, thus pre-classification and approximate methods may become recommendable.

While in this work a fixed-length noise model was used to limit the number of parameter changes, variable-length methods are equally applicable to noise, where the variation between unit lengths may be even greater than for speech.

6. CONCLUSIONS

We proposed methods for acquiring variable-length long-context speech bases for noise robust speech separation and recognition. Spectral features, state labels, and a combination of both were used for clustering speech patterns to atoms via longest-first segment search. Applied to 1st CHiME Challenge data, the methods produced speaker-adaptive bases with atom lengths ranging from 10 to 50 frames. We managed to reduce model sizes by up to 40% from already compact fixed-length bases, while achieving similar or better separation and speech recognition results. The presented methods can be used to model large vocabulary speech and non-stationary noise for better applicability to real world ASR scenarios.

7. REFERENCES

[1] O. Räsänen and U.K. Laine, "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences," *Pattern Recognition*, vol. 45, no. 1, pp. 606–616, 2012.

[2] T.M. Elliott and F.E. Frédéric, "The Modulation Transfer Function for Speech Intelligibility," *PLoS Computational Biololgy*, vol. 5, no. 3, pp. e1000302, 2009.

[3] T.N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J.F. Gemmeke, J.R. Bellegarda, and S. Sundaram, "Exemplar-Based Processing for Speech Recognition: An Overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98–113, 2012.

[4] H. Hermansky and S. Sharma, "TRAPs – Classifiers of Temporal Patterns," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998, pp. 1003–1006.

[5] H. Van hamme, "HAC-models: a Novel Approach to Continuous Speech Recognition," in *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008, pp. 2554–2557.

[6] O. Räsänen, "A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events," *Cognition*, vol. 120, no. 2, pp. 149–176, 2011.

[7] K. Kinoshita, M. Souden, M. Delcroix, and T. Nakatani, "Single Channel Dereverberation Using Example-Based Speech Enhancement with Uncertainty Decoding Technique," in *Proceedings of INTERSPEECH*, Florence, Italy, 2011, pp. 197–200.

[8] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S. Hahm, and A. Nakamura, "Speech Recognition in the Presence of Highly Non-stationary Noise Based on Spatial, Spectral and Temporal Speech/Noise Modeling Combined with Dynamic Variance Adaptation," in *Proceedings of 1st CHiME workshop*, Florence, Italy, 2011, pp. 12–17.

[9] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[10] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[11] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proceedings of 1st CHiME workshop*, Florence, Italy, 2011, pp. 24–29.

[12] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech and Language*, vol. 27, no. 3, pp. 763–779, 2013.

[13] E. Yılmaz, J.F. Gemmeke, D. Van Compernolle, and H. Van hamme, "Noise-robust Digit Recognition with Exemplar-based Sparse Representations of Variable Length," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, 2012.

[14] D. Wang and J. Tejedor, "Heterogeneous Convolutive Non-Negative Sparse Coding," in *Proceedings of INTERSPEECH*, Portland, Oregon, USA, 2012.

[15] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[16] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[17] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Compact Long Context Spectral Factorisation Models for Noise Robust Recognition of Medium Vocabulary Speech," in *Proceedings of 2nd CHiME workshop*, Vancouver, Canada, 2013, pp. 13–18.