Sparse Component Analysis for Speech Recognition in Multi-Speaker Environment

Afsaneh Asaei[1,2], *Hervé Bourlard*[1,2] *and Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
aasaei@idiap.ch, hbourlard@idiap.ch, pgarner@idiap.ch

Abstract

Sparse Component Analysis is a relatively young technique that relies upon a representation of signal occupying only a small part of a larger space. Mixtures of sparse components are disjoint in that space. As a particular application of sparsity of speech signals, we investigate the DUET blind source separation algorithm in the context of speech recognition for multi-party recordings. We show how DUET can be tuned to the particular case of speech recognition with interfering sources, and evaluate the limits of performance as the number of sources increases. We show that the separated speech fits a common metric for sparsity, and conclude that sparsity assumptions lead to good performance in speech separation and hence ought to benefit other aspects of the speech recognition chain.
Index Terms: sparse component analysis, overlapping speech, speech recognition

1. Introduction

Human listeners recognize speech even in very adverse acoustical environments with strong interfering sound sources. However, for state-of-the-art automatic speech recognition (ASR) systems, this scenario is very challenging with little success achieved. The discrepancy between human and machine performance has motivated many feature extraction approaches inspired by modeling the human auditory system. Perceptual modeling indicates that a sparse representation exists in the auditory cortex and the more accurate the auditory model, the sparser the representation [1]. The effectiveness of sparsity assumptions for speech recognition has yet to be verified. This paper gives a preliminary evaluation of exploiting sparsity assumptions for robust speech recognition. One implication of such assumptions is that it should be possible to separate the overlapping speech of two or more speakers in a sparse domain where the recovered components preserve enough information to be recognized well.

The problem of overlapping speech is one of the major challenges of speech recognition systems in multi-speaker environments and distant-talking applications. As identified in [2], around 10–15% of words or 50% of speech segments in a meeting or telephone conversation contain some degree of overlapping speech. These overlapped speech segments result in an absolute increase in speech recognition word error rate of 15–30%. Therefore, any system designed to recognize speech in multi-speaker environments is required to initially separate the speech from each individual prior to recognition. Previous approaches to tackle this problem can be grouped into three categories. The first category relies on spatial filtering techniques based on beamforming to capture a specific target by steering the beam pattern of a microphone array [3]. The second category incorporates the favorable tool of Independent Component Analysis (ICA) to identify the mixing model based on assumptions of statistical independency and non-Gaussianity. The sources are then recovered linearly by least square optimization or matrix pseudo-inversion [4]. The third category is based on sparse representation of the signal, also known as sparse component analysis (SCA). These techniques have turned out in the last few years to be a successful tool to estimate the mixing parameters and non-linear recovery of the source components [5, 6].

Previous work to evaluate the source separation approaches to perform speech recognition has been largely confined to the first two categories, and has imposed that the number of sources to be less than or equal to the number of microphones. When this condition is not satisfied, the problem is under-determined and traditional demixing approaches cannot be applied. However, given a sparse representation of the source in a transform domain, it is possible to recover the components belonging to each speaker and obtain the original signal. There is little literature on evaluating the capability of sparse techniques for speech recognition, with [7, 8] being of particular note. Previous work incorporated the sparse components in *missing data* speech recognition [7]. In this paper, we show that sparsity of speech in the time-frequency domain can be efficiently exploited in more conventional speech recognition systems. The results obtained show significant improvement to the previous missing data speech recognition approach.

The objective of the research presented in this paper is two-fold: Generally, it is about evaluating the capability of sparse techniques to allow speech recognition in overlapping conditions. More specifically, it aims to demonstrate that acknowledging sparsity leads to more robust representation of the speech signal in multi-speaker environments. This study focusses in particular on the Degenerate Unmixing Estimation Technique (DUET) to recover the sparse components of speech in the spectro-temporal domain. It is shown that these components in fact preserve the speech information to be recognized by a conventional speech recognition system.

The rest of the paper is organized as follows: The DUET source separation approach is briefly discussed in Section 2 along with contributions to address the challenges encountered while employing it to do source separation and speech recognition. Section 3 presents experimental results to evaluate the demixing performance and analyses of sparsity with a perspective to the future work. Conclusions are drawn in Section 4.

2. Source Separation in a Sparse Domain

2.1. Problem Definition

The underlying principle behind DUET is that only one source is active at any time-frequency (t-f) point. This assumption can be stated mathematically as:

$$S_j(\tau,\omega)S_k(\tau,\omega) \approx 0 \qquad \forall j \neq k, \tag{1}$$

where $S_j(\tau,\omega)$ is the windowed short-time Fourier transform (STFT) of the source j when the analysis window is centered at time τ, and ω indicates the frequency. Given that each t-f component belongs to only one of the sources, separation of these components can be achieved by applying a function which gives a unique label to the points associated with each source. Assuming that the room is anechoic, thus the mixtures are attenuated and delayed versions of the original signals along a direct path, the mixing model can be approximated as:

$$\begin{bmatrix} X_1(\tau,\omega) \\ X_2(\tau,\omega) \end{bmatrix} \approx \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-i\omega\delta_1} & \dots & a_J e^{-i\omega\delta_J} \end{bmatrix} \begin{bmatrix} S_1(\tau,\omega) \\ \dots \\ S_J(\tau,\omega) \end{bmatrix}, \tag{2}$$

where $X_1(\tau,\omega)$ and $X_2(\tau,\omega)$ is STFT of the signal captured by distant microphones. a_j and δ_j are the relative attenuation and delay parameters of the source j (proportional to the relative distance of the source to the two microphones). The total number of sources is $J \geq 2$. Based on (1) and (2), an instantaneous estimate of the mixing parameters can be obtained by applying the magnitude and phase operator onto the complex STFT ratio of the microphone signals:

$$\tilde{a} = \left| \frac{X_2(\tau,\omega)}{X_1(\tau,\omega)} \right| \quad \text{and} \quad \tilde{\delta} = -\frac{1}{\omega} \arg\left(\frac{X_1(\tau,\omega)}{X_2(\tau,\omega)} \right), \tag{3}$$

Given that each source has a unique mixing parameter (spatial signature), the problem is how to identify the actual mixing parameters from these instantaneous estimates and estimate the sources.

2.2. Estimation of the Mixing Parameters and Sources

Assuming the contribution of the interfering sources to be independent Gaussian noise and maximizing the likelihood of the mixed signals given the source (S) and mixing parameters (a and δ), a closed-form estimator is obtained [9]. We proceed from the result of [9] that states that the number of sources and their corresponding mixing parameters can be identified based on the number and location of the peaks in a 2D weighted histogram, where the $(\tilde{\alpha}, \tilde{\delta})$ pairs are used to indicate the indices into the histogram and each point is weighted by

$$|X_1(\tau,\omega)X_2(\tau,\omega)|^p \omega^q, \tag{4}$$

where $\tilde{\alpha} = \tilde{a} - 1/\tilde{a}$ is the symmetric attenuation used to obtain maximum likelihood (ML) estimate [9] and p and q are hyper-parameters chosen for various weighting schemes. In the 2D histogram constructed in this way, clusters of weights will emerge centered on the actual mixing parameter pairs corresponding to the source locations. We found that the following steps are required to achieve high recognition performance.

2.2.1. Hyper-parameter Optimization

The weighted histogram described above is based on a ML estimate for the mixing parameters. It has been shown in [10] that it is possible to optimize the hyper-parameters p, q for a generalized ML estimate and it has been suggested that $p = 1$ and $q = 0$ is a good default choice. Following a crude grid search, we find that choosing $p = 0.5$ and $q = 1$ applied on subband frequency components gives the best recognition performance.

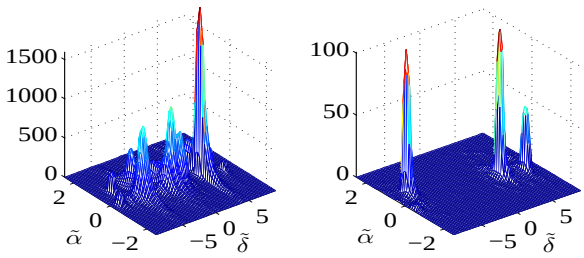

Figure 1: Left: Fullband weighted histogram, Right: Subband weighted histogram ($100 \leq \omega \leq 570$). The actual number of sources is 3, p=0.5 and q=1. Distance between microphones is 0.03m.

2.2.2. Subband Weighted Histogram

The histogram based localization approach imposes that the microphones are sufficiently close to avoid the delay estimate from the complex STFT to wrap around. This requires that

$$|\omega\delta_j| < \pi. \tag{5}$$

For the cases where this constraint is not satisfied, we define a safe-delay margin (D) based on the maximum high-frequency component and tile a number of histograms constructed from delaying one mixture against the other by products of D. The histograms are then appended to obtain a large histogram with a big delay range. To prevent spurious peaks due to phase-wrapping, we propose to consider only the subband frequency components that satisfy (5). Figure 1 illustrates a subband histogram and its fullband counterpart. As can be seen, the subband histogram contains very localized peaks around the actual mixing parameters whereas the fullband histogram has many spurious peaks that prevent accurate localization of the sources. For a speech signal, there are high energy components below 400 Hz due to pitch and the first formant frequency of the high vowels (e.g., /i/ and /u/). This corresponds to the subband histogram approach being useful for microphone separations up to 40 cm.

2.2.3. SNR-based Spectral Smoothing

Having estimated the mixing parameters as the peak centers of the histogram, they can be used to label all t-f points and construct J disjoint masks to separate the components belonging to each of the speakers. We are interested in evaluating the amount of information recovered based on disjointness assumptions for speech recognition systems. The main difficulty that we observed is that the standard feature extraction approaches for speech recognition are sensitive to the gaps in the spectra resulting from masking. Previous authors [7] have used missing data techniques to deal with these missing values. In this study we investigate a two step procedure: First we set the missing values to zero and use overlap-add (OLA) to reconstruct a time domain signal. Then, we add white Gaussian noise to the signal within a specific signal to noise ratio (SNR). By preventing the effect of zeros on the feature extraction, this will lead to smooth the spectral shape at the discontinuities. OLA is also a convenient mean of changing the DFT size and period.

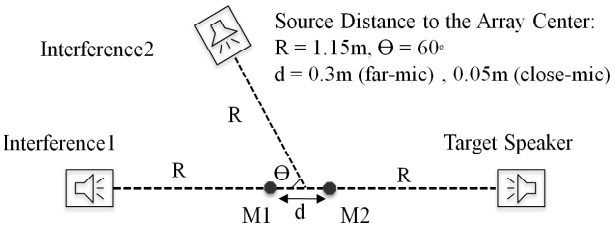

Figure 2: Overhead view of the room set-up

3. Experiments

We conducted experiments to try to answer the following questions:

1. Are the sparse source separation assumptions valid while incorporated for recognition of overlapping speech? i.e., Does DUET work well with conventional speech recognition systems?
2. What is the limit of performance? i.e., How far we can push the disjointness/sparsity assumptions?
3. Given the spectro-temporal representation of speech signal, is the salient information needed for speech recognition preserved only in a small fraction of the whole components? i.e., How well does it fit a common metric for sparsity?

3.1. Database

The experiments are all performed in the framework of Aurora 2 [11]. This database is designed to evaluate the performance of speech recognition algorithms in noisy conditions. A fixed HTK back-end was trained on multi-condition data with different noise types including those of Subway, Babble, Car and Exhibition at 5 SNR levels as well as clean data. Overlapping speech was synthesized by mixing clean Aurora 2 test utterances with interfering sentences from the HTIMIT database. To make sure that the results are generalizable for any interferences, we have randomly chosen a pool of 40 sentences balanced among male and female utterances from HTIMIT. For each test sample, interferences are randomly chosen out of this subset to construct two mixtures. All files are normalized prior to mixing and to compensate for the difference between the file lengths, the interferences are looped. Figure 2 shows an overhead view of the physical set-up being simulated.

3.2. Simulation Parameters

The subband histogram is constructed for the frequency band of 100–570 Hz. The lower band is chosen based on a lowest notional pitch frequency. Although, HTIMIT and AURORA are both telephony speech, the frequency components below 300 Hz are not completely suppressed. Recall from equation (3) that the ratio of the two mixtures is used for instantaneous mixing parameter estimates $(\tilde{\alpha}, \tilde{\delta})$; thus the channel response which is the same for both mixtures is canceled. The upper-band is chosen to satisfy equation (5) and prevent phase-wrapping. The histogram resolutions in attenuation and delay are 0.06 and 0.14 samples respectively. The histogram attenuation width is $|\tilde{\alpha}| \leq 2.5$. For the close-microphone scenario, the histogram delay width is $|\tilde{\delta}| \leq 4$. In the far-microphone case, 3 histograms are appended together, each obtained by delaying the second mixture by 4 and -4 samples. Therefore, the delay-width of the big histogram is $|\tilde{\delta}| \leq 8$ samples. The target is detected based on the geometric proximity to the position of interest. Notice that the sub-band histogram is only used to estimate the mixing parameters (roughly speaking, source localization) whereas the separation of source components are all performed for the whole frequency band. The analysis and synthesis window for source separation and signal reconstruction is Hann to facilitate OLA. The size of the window is 125 ms. Following a crude grid search, we find that choosing $p = 0.5$ and $q = 1$ and adding white Gaussian noise at 37 dB to the demixed signal gives the best performance.

The separated signal is then presented to the standard Aurora 2 speech recognition system. The speech signal is processed in blocks of 25 ms with a shift of 10 ms to extract 13 MFCC cepstral coefficients per frame. These coefficients after cepstral mean/variance normalization are appended to their delta and delta-delta derivatives to obtain a 39 dimensional feature vector for every 10 ms of speech.

3.3. Results and Analyses

3.3.1. Speech Recognition Performance

Table 1 gives the recognition results of the demixed speech for the clean and multi-condition training.

Table 1: Word accuracy for mixtures and separated components smoothed with OLA and adding white Gaussian noise (WGN).

Mic d(cm)	Train Cond.	Aurora(%) Baseline	DUET(%) OLA	DUET(%) OLA+WGN
5	Clean	50.09	77.03	79.29
	MultiCon.	39.62	81.09	91.14
30	Clean	51.35	80.31	86.34
	Multi-Con.	44	79.95	**93.35**

The recognition accuracy shows the significant potential of disjointness assumptions in the spectro-temporal domain to demix the signals while preserving the salient information to perform speech recognition. The best results are obtained for far-microphones. The histogram peaks for far-microphones are well localized, and the peak regions are clearly distinct, whereas the peak regions in the close-microphone histogram have some degree of overlapping. Experiments on full-band weighted histogram as well as non-frequency weighted histogram ($q = 0$) yielded consistently poor results, more than 10% reduction in recognition rate. For the close-microphone scenario, the proposed sub-band weighting scheme is still the best choice. Furthermore, we observed that adding a negligible amount of Gaussian noise (SNR=37dB) improves the recognition results up to 14%. The improvement is obtained for both clean and multi-condition training. We could justify this by considering that adding noise specially improves the non-voiced speech which is mainly potential non-disjoint/sparse part of speech signal with an inherent noisy nature.

3.3.2. Performance Limit

To approach the question on how far we can push the disjointness/sparsity assumptions, we set up some experiments to do recognition of the separated speech while the number of interferences is increased up to 10. We intend to quantify the sparsity of informative coefficients of speech signal in terms of recognition rate. The set of mixing parameters are chosen to provide a different spatial signature for each speaker. The recognition rate after separation is depicted in Figure 3. We can conclude from this observation that the salient information needed to recognize

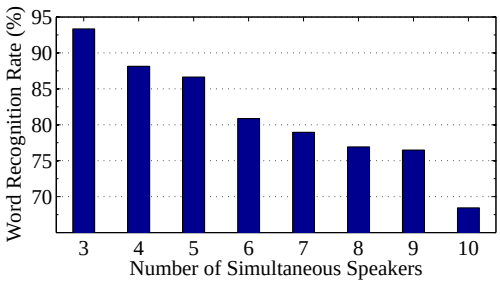

Figure 3: Spectro-temporal disjointness of overlapping speech quantified in terms of ASR performance

speech is in fact in a small fraction of the t-f coefficients and it is quite unlikely for the speech interferences to overlap these sparse structures. Furthermore, we observed that increasing the number of interferences degrades the accuracy of the separation by attenuation and the histogram clusters are separated mainly due to the different delay parameters. This observation emphasizes that the sparse source separation techniques which exploit both the delay and attenuation has potential to exhibit more robustness.

3.3.3. Sparsity of Speech in Spectro-Temporal Domain

The results of the experiments for the increased number of interferences were very intriguing. To examine how it fits a common metric for sparsity, we investigate the spectro-temporal components recovered by DUET. For the natural signals to be closely approximated as sparse, their coefficients ζ must have a rapid power-law decay when sorted [12]

$$|\zeta_{(i)}| \leq \gamma i^{\frac{-1}{r}} \qquad r \leq 1, \tag{6}$$

where $\zeta_{(i)}, i = 1, 2, ...$ denotes the coefficients of ζ when sorted from largest to smallest. Plotting the sorted absolute value of the recovered t-f components vs. their index is illustrated in figure (4) which satisfies equation (6) and exhibits power-law decay, though more than 200 coefficients are needed to get into a regime where the coefficient decay is better than -1. Based on this observation, the speech representation in spectro-temporal space can be approximated to be sparse. This observation has been already been shown to be beneficial for speech recognition and verified concisely through equation (6). This motivates investigating the sparse features in an integrated framework for source separation and speech recognition. As proposed by [1], a dictionary of Gabor atoms can be learned for ASR where the coefficients of the decomposition of speech is directly applicable as ASR features. Interestingly, these features also exhibit sparsity and satisfy equation (6). Thus, they have potential to be exploited in an ASR front-end which is robust to overlapping.

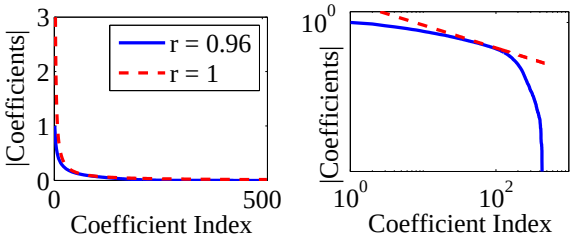

Figure 4: Power-law decay of the recovered t-f components depicted in linear (left) and logarithmic (right) axes.

4. Conclusions

We have evaluated the sparsity assumptions incorporated in sparse component analysis in the framework of DUET for speech recognition in a multi-speaker environment. Recognition results after demixing show that the salient information needed to recognize speech is in a fraction of disjoint t-f components. Setting the rest of the coefficients to zero and smoothing the spectral discontinuities by OLA and adding white Gaussian noise, the demixed signal can be recognised using a conventional speech recognition system. These analyses strengthen the benefit of sparse assumptions for recognition of overlapping speech. When the speech signal is represented in a sparse domain, separation of the components becomes straightforward, and these components in fact preserve the information to be recognized well. This motivates more research on identifying a domain/dictionary where the speech representation/decomposition is sparse and these coefficients are directly applicable for speech recognition.

5. Acknowledgements

Authors would like to thank Prof. Volkan Cevher for the fruitful discussions and valuable comments.
The research leading to these results has received funding from the European Union under the Marie-Curie Training project SCALE (Speech Communication with Adaptive LEarning), FP7 grant agreement number 213850.

6. References

[1] Kleinschmidt, M., “Robust Speech Recognition Based on Spectrotemporal Processing”, PhD thesis, Univ. Oldenburg, 2002.

[2] Shriberg, E., Stolcke A. and Baron, D., “Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation”, In Proceedings of Eurospeech 2001.

[3] Bourgeois, J. and Minker, W., “Time-Domain Beamforming and Blind Source Separation: Speech Input in the Car Environment”, Lecture Notes in Electrical Engineering, Springer 2009.

[4] Chien, J.T. and Chen, B.C., “A New Independent Component Analysis for Speech Recognition and Separation”, IEEE transactions on audio, speech, and language processing, vol. 14, July 2006.

[5] Zibulevsky, M. and Pearlmutter, B. A., “Blind Source Separation by Sparse Decomposition in a Signal Dictionary”, Neural Computation, 13(4):863-882, 2001.

[6] Gribnoval, R., “Sparse Decomposition of Stereo Signals with Matching Pursuit and Application to Blind Separation of More than two Sources from a Stereo Mixture”, ICASSP 2002.

[7] Kuhne, M., Togneri R. and Nordholm, S., “Mel-Spectrographic Mask Estimation for Missing Data Speech Recognition using Short-Time-Fourier-Transform Ratio Estimators”, ICASSP 2007.

[8] Kuhne, M., Togneri, R. and Nordholm, S., “Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition”, in Speech Recognition, Technologies and Applications. I-Tech, 2008

[9] Yilmaz, O. and Rickard S., “Blind Separation of Speech Mixtures via Time-Frequency Masking”, IEEE Transactions on Signal Processing, vol. 52, pp. 1830-1847, 2004.

[10] Makino, S., Lee T. and Hiroshi, S., (Eds.) “Blind Speech Separation”, Chapter 8: The DUET Blind Source Separation Algorithm”, Springer Netherlands, 2007.

[11] Pearce, D. and Hirsch, H., “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions”, ICSLP 2000.

[12] Cevher V., “Learning with Compressible Priors,“ NIPS, Vancouver, B.C., Canada, 7-12 December 2009.