

Single Channel Speech Separation Based on Discriminative Dictionary

Dr. Rajeev Shrivastava¹, Mohammad Javeed², Anupendra Singh³

¹Professor, Department of ECE, Sree Dattha Institute of Engineering & Science, Hyderabad, India

²Assistant Professor, Department of ECE, Sree Dattha Institute of Engineering & Science, Hyderabad, India

³Lecturer, Department of ETE, Government Women's Polytechnic College Jabalpur, India

ABSTRACT

This project concentrates on single channel separation (SCSS) based on dictionary learning method (DL) to improve the performance of the sparsity of SCSS. The conventional approaches consider the sub words as independent entities and learn separately on the short-term Fourier transformer (STFT) domain using their corresponding training set. However, we take into account the relationship between the sub-dictionaries and optimize the sub-dictionaries jointly in the time domain. By fulfilling a designed discrimination limitation, a structured dictionary, whose atoms have better correspondences to speaker labels, is learned so that the sources can be recovered by the corresponding reconstruction after sparse coding. An algorithm, consisting of sparse coding phase and dictionary update phase, is proposed to deal with this DL optimization problem. Two strategies, i.e. direct learning and adaptive learning, are presented to select the educational sets used to learn the discriminatory dictionary. The results have been done in Xilinx ISE tool and coding performed with verilog language and have shown that sufficient results compared to the existing system.

Keyword: - AR, SCSS, Fourier Transform, Mixed signal design.

1. INTRODUCTION

SPEECH separation [1]–[3] has been a topic of research in the last few decades, not only because of its difficulty but also for several potential applications. For example, in noisy environments, a robust speech separation is often required as the pre-processing stage prior to the target application, such as hearing aids, automatic speech recognition and speech coding. A single-channel speech separation (SCSS) system [4]–[14], which is applicable to the case that only one microphone is available, aims at recovering the underlying speakers' signals from a single mixed signal [15]. The CASA approaches seek discriminative features in the mixed signal to separate the speech signals. In contrast, the model-based approaches mainly rely on a priori knowledge of sources obtained during a training phase. CASA use multi-pitch estimation methods to extract pitch estimates of the speakers directly from the mixture. The separation performance of CASA-based methods, as a consequence, is predominantly affected by the accuracy of the multi-pitch estimator, especially when the pitch of one of the speakers is masked by the others. Model-based methods use pre-trained speaker models as a priori information to constrain the solution of the ill-conditioned SCSS problem. In particular, source-specific speaker models are incorporated to capture specific characteristics of individual speakers at each frame. Non-negative matrix factorization (NMF) [5], [16], [17] and complex matrix factorization (CMF) [8], [15], [18] based SCSS are the typical model-based methods. Many existing model-based approaches are performed in the short-time Fourier transform (STFT) domain, while our proposed approaches are performed in the time domain and use the concept of dictionary learning (DL). It should be noted that the existing approaches [5], [8], [15] generally use the speakers' respective speech signals separately to capture specific features and ignore the fact that the discrimination comes not only from the object speaker but also from the interference speaker.

Speech technology works reasonably in matched conditions but rapidly degrades when there is acoustic mismatch between the training and test conditions. Although multi-condition training can improve the performance [16], realistic scenarios can benefit from more robustness without requiring training data from the target acoustic environment. In this paper, we develop a feature extraction scheme which attempts to address robustness in noisy and reverberant environments. Automatic speech recognition technology has a high potential for improving the learning experience of students in an educational setting. Some of the key theoretical areas involved in developing automatic speech recognition systems for educational use; namely the applications of the technology in education, prominent feature extraction and noise cancellation techniques used with audio speech data as well as some of the recent neural network based machine learning models capable of keyword spotting or continuous speech recognition

shown in [7]. Authors have discussed some of the traditional feature extraction techniques that are commonly used in the areas of language identification speech recognition, and speaker verification, and their pros and cons in [18]. Due to the nonlinear nature of speech, LPC are not generally used for speech estimation. It was discussed that the most frequently used feature extraction techniques are MFCC, LPC and PLP in the areas of speaker verification and speech recognition applications but recently hybrid features are overcoming the traditional features [3].

The block schematic for the proposed feature extraction is shown in Fig. 1. Long segments of the input speech signal (10s of non-overlapping windows) are transformed using a discrete cosine transform . The full-band DCT signal is windowed into a set of 96 over-lapping linear sub-bands in the frequency range of 125-3700 Hz. In each sub-band, linear prediction is applied on the sub-band DCT components to estimate an all-pole representation of Hilbert envelope. This constitutes the temporal AR modeling stage. The FDLP envelopes from various sub-bands are stacked together to obtain a two-dimensional representation as shown in Fig. 1. (25ms with a shift of 10ms)

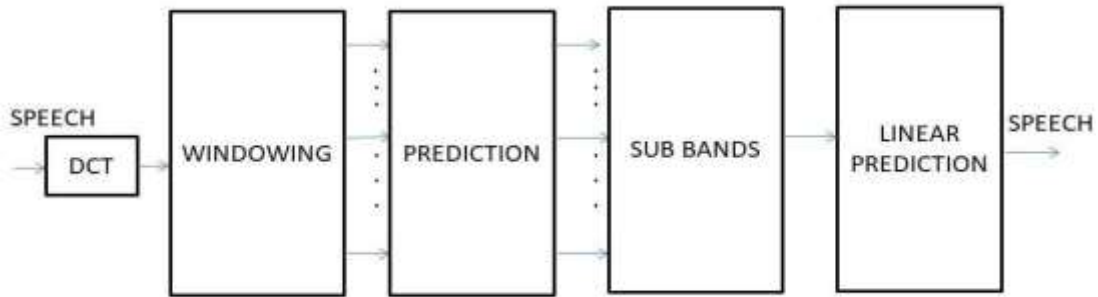


Figure 1: Block schematic of the proposed feature extraction using spectro-temporal AR Models.

The output of the integration process provides an estimate of the power spectrum of signal in the short-term frame level. The frequency resolution of this power spectrum is equal to the initial sub-band decomposition of 96 bands. These power spectral estimates are transformed to temporal autocorrelation estimates using inverse Fourier transform and the resulting autocorrelation sequence is used for time domain linear prediction (TDLP). We derive 13 campestral coefficients from the all-pole approximation of the 96 point short term power spectrum. The delta and acceleration coefficients are extracted to obtain 39 dimensional features. The SCSS solves the following problem: recovering N underlying speech signals $S_i, i=1,2,\dots, N$ from the mixed speech signal.

$$\mathbf{x}(t) = \sum_i \mathbf{s}_i(t), 1 \leq t \leq T.$$

For convenience of description, the case of 2 source speech signals is considered, i.e.,

$$\mathbf{x} = \mathbf{s}_1 + \mathbf{s}_2.$$

It should be noted that our approaches, which will be described below, can be easily extended to the general case when the number of source speech signals is more than 2. We carry out separation frame by frame with the window length of and 50% overlap. We define some notations here: $D=[D_1,D_2]$, denotes the learned structured dictionary which is comprised of two sub-dictionaries D_1,D_2 . The two approachable methods direct learning and adaptive learning methods have been shown in figure 2.

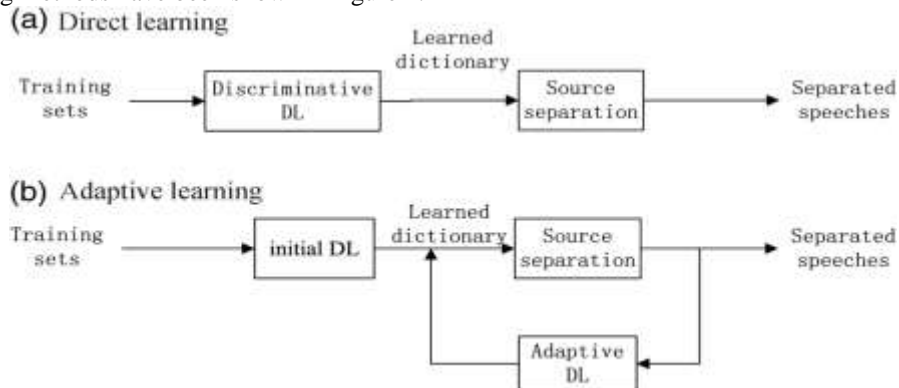


Figure 2: (a) Direct learning method (b) Adaptive learning method

II METHODOLOGY

A. Rate-Scale filtering using AR Models

A temporal modulation filter is referred to as a rate filter and a spectral modulation filter is referred to as a scale filter. In the proposed feature extraction framework, the AR modeling process represents a filter impulse response, whose frequency response (“time response” in the case of the temporal AR filter) can be controlled by the model order. A lower model order represents more smoothing in a given domain, while the higher model captures finer details. Thus, various streams of spectrographic representations can be generated from the proposed framework using different choices of model order for temporal and spectral AR models as shown in Fig. 2. The low-rate lowscale representations represent broad energy variations in the signal as seen in Fig. 2 (b). The other configuration using higher order for the AR models is shown in Fig. 2 (c) where more details about the various events in the spectrogram are evident. A higher order could also mean that such AR models may carry information about noise or reverberation artifacts that is present in the finer details of the spectrogram in its spectral or temporal directions. In Sec. 4, we provide some experiments showing the effect of model order on the speaker recognition performance.

A. Robustness to Noise

When a speech signal is corrupted with noise or reverberation, the valleys in the sub-band envelopes are dominated by noise. Even with moderate amounts of distortion, the low-energy regions are substantially modified and cause acoustic mismatch with the clean training data. Since the AR modeling tends to fit the high energy regions with good accuracy. This is illustrated in Fig. 2 where we plot a portion of clean speech signal, speech with additive noise (babble noise at 10 dB SNR) and speech with artificial reverberation (reverberation time of 300

ms). The spectrographic representation obtained from mel frequency representation is shown in the second panel and the corresponding representation obtained from spectro-temporal AR models is shown in the bottom panel. In comparison with the mel spectrogram, the representation obtained from AR modeling emphasizes the high energy regions. Thus, such a representation can be more similar for the clean and the noisy versions of the same signal. This is desirable and contributes to improved robustness when these features are used for speaker recognition in noisy environments.

III RESULTS

Figure 3 shows the timing details of the adaptive learning method. Figure 6 shows the details of timing of direct learning method. Figures 4 and 5 shows the RTL view of the adaptive learning methods. Table I shows the timing difference between the two methods. From the experimental results we can say that the adaptive learning method is having the better performance than the direct learning method. Implementation have done in Xilinx ISE 10.1 version. Code written in the Verilog Hardware description language.

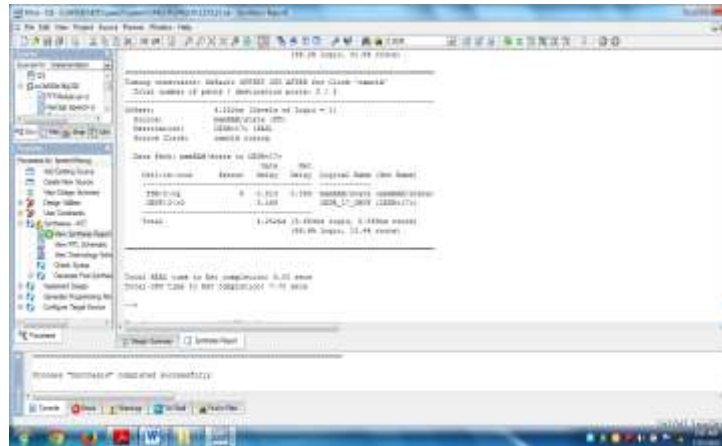


Figure 3 timing details of the design.

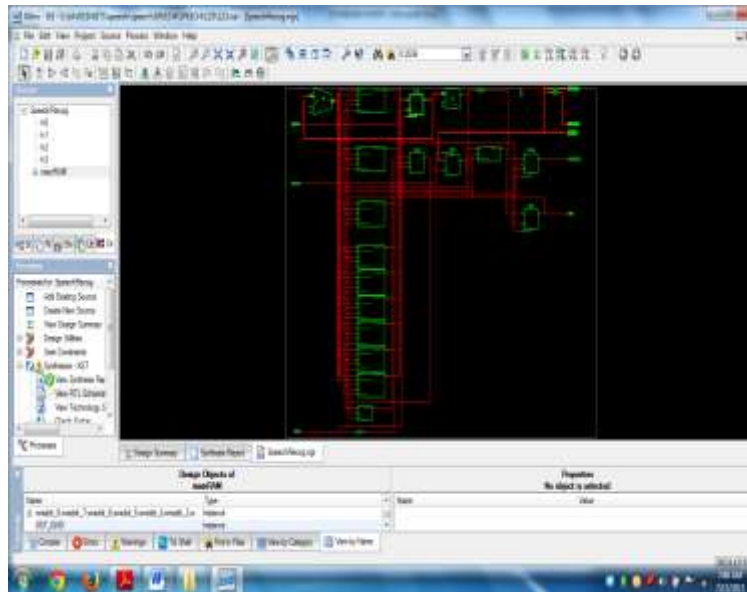


Figure 4 RTL view of the design

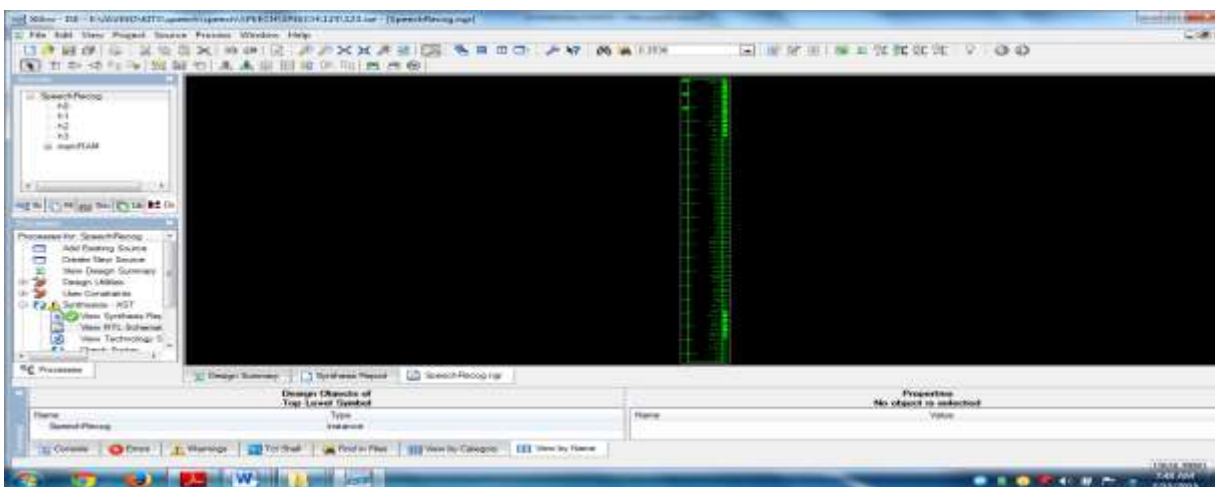


Figure 5 RTL view of the design

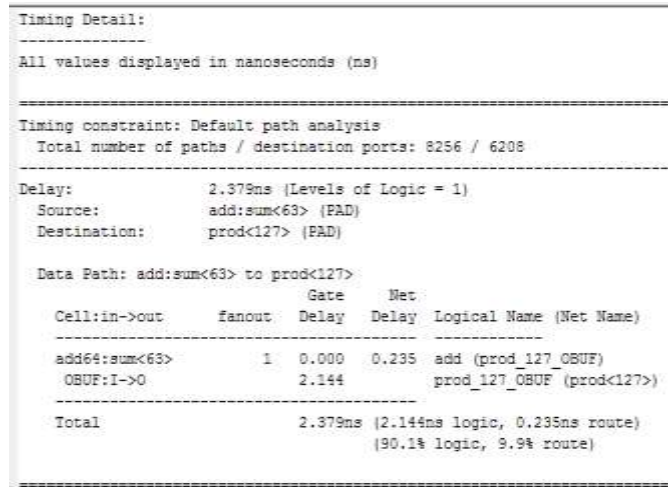


Figure 6 Timing details of the Adaptive learning method

TABLE I. TIMING TABLE

	DLM	ADLM
TIME (ns)	4.134ns	2.379ns

IV. CONCLUSION

We have presented a novel DL method (DDL) to improve the SCSS performance in this paper. Unlike the conventional methods, we take not only the object speaker but also the interference speaker into account and the sub-dictionaries are optimized jointly in the time domain. The learned discriminative dictionary has the property that its atoms have better correspondences to the speaker labels so that the sources can be recovered by the corresponding reconstruction after sparse coding. In order to deal with this DL optimization problem, an algorithm, consisting of sparse coding stage and dictionary updating stage, has been proposed. These two stages are iteratively performed for a specified number of times or until convergence. In addition, two strategies have been proposed to select the training sets: direct learning and adaptive learning. The adaptive learning approach produces better performance than the direct learning approach at the expense of more time cost.

5. REFERENCES

[1] A. Jourjine, S. Rickard, and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2000, pp. 2985–2988.

[2] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[3] G. Bao, Z. Ye, X. Xu, and Y. Zhou, "A compressed sensing approach to blind separation of speech mixture based on a two-layer sparsity model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 899–906, May 2013.

[4] S. T. Roweis, "One microphone source separation," in *NIPS*, 2000, pp. 793–799.

[5] M. Schmidt and R. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. Interspeech*, 2006, pp. 2614–2617.




[6] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[7] Y. K. Lee, I. S. Lee, and O. W. Kwon, "Single-channel speech separation using phase-based methods," *IEEE Trans. Consumer Electron.*, vol. 56, no. 4, pp. 2453–2459, Nov. 2010.

[8] B. J. King and L. Atlas, "Single-channel source separation using complex matrix factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 8, pp. 2591–2597, Nov. 2011.

- [9] B.Gao, W. L. Woo, and S. S. Dlay, "Single-channel source separation using EMD-subband variable regularized sparse features," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 961–976, May 2011.
- [10] P. Mowlaee, R. Saeidi, M. G. Christensen, Z. H. Tan, T. Kinnunen, P. Franti, and S. H. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 9, pp. 2586–2601, Nov. 2012.
- [11] C. Demir, M. Saraclar, and A. Cemgil, "Single-channel speech-music separation for robust ASR with mixture models," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 4, pp. 725–736, Apr. 2013.
- [12] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 2014–2023, Nov. 2006.
- [13] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, 1994.
- [14] D. P. Ellis, "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures," *Speech Commun.*, vol. 27, no. 3, pp. 281–298, 1999.
- [15] B. J. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, 2010, pp. 4206–4209.

BIOGRAPHIES (Not Essential)

	<p>Dr. RAJEEV SHRIVASTAVA currently working as a Professor in department of ECE in Sree Dattha Institute of Engineering and Science, Hyderabad. His Ph.D. is from JVV University and M.Tech, B.Tech are from RGPV, BHOPAL. He has published more than 63 International research articles in reputed journals and conferences. He has given his Key note speeches and research presentations in reputed conferences in India as well as abroad countries like Malaysia, Thailand etc. He is an editorial board member, Advisory board member and reviewer for more than 10 international Journals. He has 5 patents to his ideas and 3 International level Research awards which includes young scientist, excellent research awards. He is an author of 3 books and 3 book chapters.</p>
	<p>Dr. MD.JAVEED currently working as an Assistant Professor and Project Coordinator for the department of ECE in Sree Dattha Institute of Engineering and Science, Hyderabad. His Ph.D. is from Mewar University and M.Tech, B.Tech are from JNTU, Hyderabad. He has published more than 40 International research articles in reputed journals and conferences. He has given his Key note speeches and research presentations in reputed conferences in India as well as abroad countries like Malaysia, Thailand etc. He is an editorial board member, Advisory board member and reviewer for more than 10 international Journals. He has 5 patents to his ideas and 8 International level Research awards which includes young scientist, excellent research awards. As a project coordinator he has guided and developed hundreds of projects in his 7 years of experience for students as well as for product development. He is an author of 2 books and two book chapters.</p>
	<p>I did my engineering In Electronics and telecommunications from RGPV Bhopal and then joined as Assistant professor in GRKIST Jabalpur. Worked as Assistant professor in GRKIST Jabalpur for nearly five years. Then joined Government women's polytechnic college jabalpur as LECTURER in ETE department in 2010. Working there since February 2010. Completed M.tech in digital communication from RGPV BHOPAL IN 2019. Key areas of interest include antenna, image processing, vlsi and RADAR.</p>