# Achieving Robust Automatic Speech Recognition and Exemplar Enhanced Speech Approach by Coupled Dictionaries for Improved Signal-To-Distortion Ratio

KAKARLA ASHOKKUMAR,Email.Id: ashokkumar2992@gmail.com

Lakshma Reddy, M.tech, Assistant Professor,Email.Id: lakshmareddy@newton.edu.in

Newton's Institute of Engineering, Alugurajupalli (V), Koppunoor Macherla(M), Guntur(D),Pin: 522426

## Abstract

*Innovative inventions related to Electro-Magnetic (EM) spectrum have created huge impact on modern world. EM spectrum comprises of three prominent domains namely (i) Digital Signal Processing, (ii) Digital Image Processing and finally (iii) Digital Speech Processing. Speech processing and its associated applications suffer from noise and distortion in speech recognition process in various high end applications. Composed speech signal processing has poses challenges and takes more time to accomplish the task. Decomposition of speech signal evaluates the unique speech coefficients statistics and noisy speech coefficients statistics in reliable way which helps to process speech signal in efficient way. In this paper, exemplar based speech enhancement system is introduced to obtain the decomposition, exemplars sampled in lower dimensional spaces are preferred over the full-resolution frequency domain for their reduced computational complexity and the ability to better generalize to unseen cases. an efficient way to directly compute the full-resolution frequency estimates of speech and noise using coupled dictionaries: an input dictionary containing atoms from the desired exemplar space to obtain the decomposition and a coupled output dictionary containing exemplars from the full resolution frequency domain. The proposed system was evaluated for various choices of input exemplars and yielded improved speech enhancement performances on the AURORA-2 and AURORA-4 databases. We further show that the proposed approach also results in improved word error rates (WERs) for the speech recognition tasks.*

**Keywords:** Speech recognition, Speech enhancement, Exemplars based system, coupled dictionaries

## 1. INTRODUCTION

Speech recordings taken from realistic environments may contain added degradations along with the required speech signal which reduces its intelligibility as well as results in poor performance in speech processing tasks like automatic speech recognition (ASR), speaker recognition, hearing aids etc. The degradation can be introduced by additive background noise, reverberation, etc., and current state-of-the-art systems employ some mechanism to suppress these artifacts to enhance the speech signal for better performance and/or intelligibility.

What does "improving speech quality" mean? It is very hard and complex to explain, however, it can be summarized as the improvement in intelligibility, and, overall, perceptual clarity and pleasantness of

the degraded speech signal. Speech enhancement and noise reduction aim to do this: improve the speech quality of a noisy signal by removing the background noises with a wide variety of techniques. Over the last years, this subject has been an important research topic due to the fact that it is required in many situations in daily life. Therefore, considering the complex characteristics of speech and the large amount of restrictions, it gets even more complicated to satisfy all objectives at once. Traditional noise reduction methods, such as Spectral Subtraction and Wiener filtering, are based on strong stationary assumption and do not works satisfactorily in the presence of real non-stationary background noise.

For evaluation, we chose two traditional exemplar-based systems as baselines; the first one which uses full-scale DFT as features, and the second which uses the Mel-integrated magnitude spectra, called the Mel features, which results in a Wiener filter with reduced degree-of-freedom. Coupled dictionaries with non-negative representation have been used to increase the spectro-temporal resolution. Here it is used to map low-dimensional spectro-temporal representations to spectral representations with sufficient frequency resolution. The simulation results obtained on the AURORA-2 database revealed that the proposed system with the Mel features as front-end results in better SDRs when compared to both the baseline systems. The paper also investigates the use of coupled dictionaries for modulation spectrogram (MS) features which has recently been successfully used for blind source separation. The proposed system with MS features also yields improved SDRs over the baseline systems.

## 2. METHOD

### 2.1. Compositional model for noisy speech using NMF

In the supervised setting, the exemplars for speech and noise are stored as columns in the dictionary matrices As and An, respectively. The exemplars may span multiple frames, T to capture temporal dynamics and are reshaped to a vector. The representation for the noisy utterance in the exemplar space, Ψ, is also obtained in the same manner by reshaping overlapping windows of length T, which is then decomposed using NMF to get the activations, X, as:

$$\psi \approx [A_s \quad A_n]\begin{bmatrix} X_s \\ X_n \end{bmatrix} = AX \quad s.t. \;\; X \geq 0 \qquad (1)$$

The approximation is done to minimize the KULLBACK - LEIBLER divergence between Ψ and AX with additional sparsity constraint on X. The frame-wise speech and noise estimates, $\hat{s}$ and $\hat{n}$ are obtained after removing the windowing effect by adding the components belonging to overlapping windows from the estimates $A_s X_s$ and $A_n X_n$ respectively. The frame-level Wiener filter in the exemplar domain is then obtained as, $W = \hat{s} \oslash (\hat{s} + \hat{n})$, where $\oslash$ denotes the element-wise division.

### 2.2. Proposed method using coupled dictionaries

In the proposed system, the NMF-based decomposition is obtained in any additive and non-negative feature space of choice which serves as the front-end of the speech enhancement system. For simplicity, the front-end features are referred to as the input features and the dictionary used to obtain the NMF compositional model is denoted as the input dictionary, $A^{in}=[A_s^{in}, A_s^{in}]$. The coupled DFT

dictionary, $A^{dft}$serves as the output dictionary with which the speech and noise estimates are directly obtained in the DFT space using the activations obtained from the front-end, $X^{in}$ as $[A_s^{dft}, X_s^{in}]$in and $[A_n^{dft}, X_n^{in}]$in respectively.

The proposed system using coupled dictionaries is summarized in Fig. 1. To obtain the dictionaries, each of the coupled exemplars for the input and the DFT dictionaries are extracted from the same piece of training data which span multiple frames of length T, followed by reshaping to form a vector. This will result in speech and noise dictionaries each for the input and DFT exemplar representations which are denoted as $A_s^{in}$, $A_n^{in}$, $A_s^{dft}$ and $A_n^{dft}$ respectively. The notations used to explain the test phase are: $\psi_{in}$ for the noisy speech represented in the input exemplar domain and $[Y]^*$ denotes the matrix obtained after removing the effect of overlapping windows in the windowed observation Y. All matrix divisions should be considered element-wise.

The proposed method thus can exploit the ability of various feature representations to separate speech from noise and can generate a Wiener filter which has full degree-of-freedom in the DFT space. In this paper, we investigate the use of the proposed approach for various features which will be discussed next.

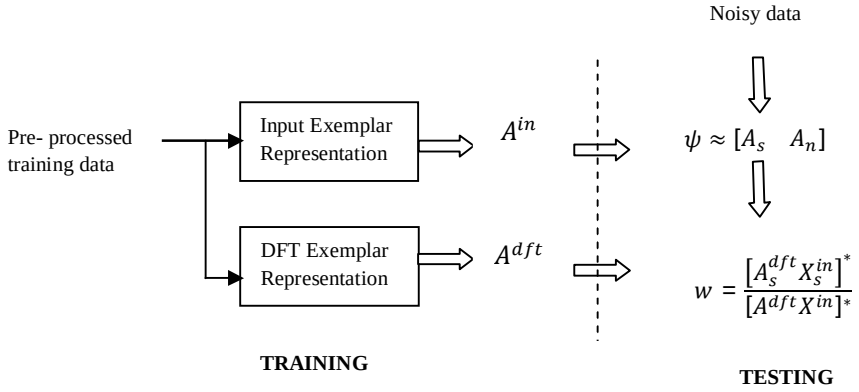


Figure 1: Fig. 1. Block diagram overview of the proposed system using coupled dictionaries

## 3. SYSTEM DESCRIPTION

### 3.1. Mel and DFT baselines

For a fair evaluation, we used two baseline systems for speech enhancement; one where the exemplars are represented in the DFT domain and the second which uses the Mel-integrated spectra for which a conversion is needed to obtain the Wiener filter on the DFT resolution. The DFT exemplars consist of full-resolution magnitude spectra with K bins per frame and segments of T frames are reshaped to get $(k.T)$ dimensional exemplars. The Mel exemplars are obtained by multiplying the DFT segments of size $k \times T$ using the FFT-to-Mel matrix, M, which contains the magnitude response of B Mel bands along its rows, followed by reshaping to vectors of size $(B.T)$. The speech and noise exemplars thus generated are stored as $A_s^{dft}$, $A_s^{mel}$, $A_n^{dft}$and $A_n^{mel}$ respectively for DFT and Mel based systems.

The DFT baseline (DFT BL) results are then obtained after finding the compositional model for noisy speech in the full-resolution DFT domain using the DFT dictionary $A^{dft} = [A_s^{dft} A_n^{dft}]$. The Wiener

filter is directly obtained in the DFT domain using the procedure explained in Section 2.1 and is then used to enhance the noisy speech [10].

To obtain the Mel baseline (Mel BL), the speech and noise estimates, $\hat{s}^*$ and $\hat{n}^*$, are first found in the Mel domain using the steps described in Section 2.1 with the Mel dictionary, $A_{mel} = [A_s^{mel}\ A_n^{mel}]$. These estimates are then extrapolated from the B dimensional Mel space to the K dimensional DFT domain using the transpose of the DFT-to-Mel matrix and the corresponding Wiener filter is then obtained, using element-wise division, as [11] :

$$W^* = \frac{M^T \hat{s}^*}{M^T \hat{s}^* + M^T \hat{n}^*} \qquad (2)$$

Since M contains triangular shaped filter-banks, this extrapolation is the same as the piece-wise linear interpolation between B points (the Mel filter-bank central frequencies) spread across the 1 to K frequency bins. The resulting filters always fall in the B-dimensional subspace defined by the columns of $M^T$ which cannot account for all the added noise content along the K dimensional DFT space. The enhanced speech obtained after applying this filter on the noisy DFT thus will result in a sub-optimal noise suppression.

### 3.2. Proposed system with Mel features

The motivation for using Mel features as the input features for the proposed system are: 1. Poor separation capability of DFT Exemplars: In the DFT based system, it has been noticed that many of the speech exemplars are activated for babble noise because of the similarity between the babble noise and speech exemplars, which in turn results in a Wiener filter which retains most of the babble noise content (ref. Fig. 2b). Similar situations were observed for unseen noise cases (ref. Fig. 2c) because the DFT exemplars lead to accurate representation of training noise cases which results in poor modeling of unseen noise cases. As a result, NMF will pick speech exemplars also to model the unseen noise content which results in poorer noise suppression.

On the other hand, Mel exemplars are found to be better able to differentiate speech from the babble noise and result in better separation (ref. Fig. 2b). The Mel features also have much lower dimensionality when compared to the DFT exemplars and reduces the risk of over fitting to seen noise cases.

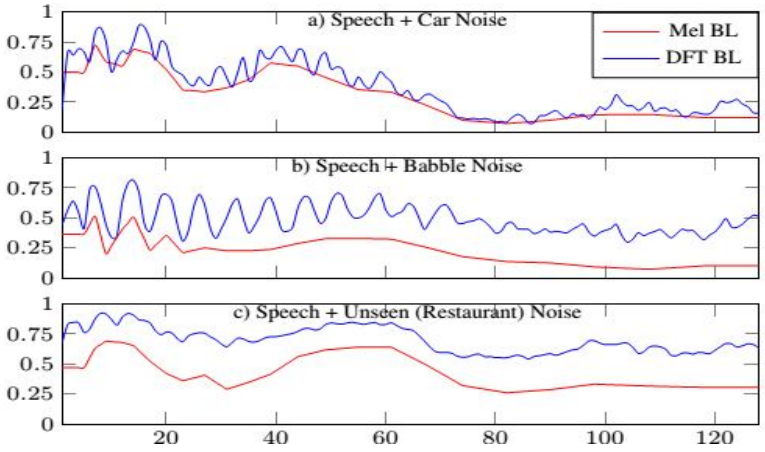

Figure 2: Filter coefficients obtained for an arbitrary frame containing speech with SNR 0dB as a function of the frequency bins. The color coding is the same for all the figures. a) For car noise which is present in the training set, the DFT baseline filter better captures the formant peaks and valleys when compared to that of the Mel baseline. b) For babble noise, speech exemplars are also activated to model the noise which results in poorer denoising. c) For the unseen restaurant noise, due to poorer modeling, the DFT exemplars result in retaining most of the noise content whereas the filter coefficients of the Mel baseline are quite smaller and result in better noise suppression.

**2. Piece-wise linear approximation of filter coefficients**

As discussed before, even though the Mel exemplars lead to better separation, the low-rank approximation of the coefficients in the DFT domain fails to capture the detailed structure of the underlying speech which can be seen in Fig. 2a. It was observed that both DFT and Mel exemplars yield almost the same separation after NMF, but with the latter resulting in a filter with lesser peaks and valleys, which is essential to capture the formant positions and pitch, yields smaller SDRs. For evaluation, the Mel and the coupled DFT dictionaries are jointly extracted first. The noisy data is converted to the Mel exemplar representation and is then decomposed using NMF with the Mel dictionary. The speech and noise estimates are then obtained directly in the DFT domain using the coupled DFT dictionary as shown in Fig. 1. This system is referred to as the Coupled Mel system.

**3.3. Proposed system with MS features**

The MS representation was proposed as part of a computational model for human hearing which relies on the low frequency amplitude modulations within various frequency bands. The MS representation for acoustical data is obtained using the procedure explained. For the NMF based system, this 3D representation of size b×T ×B, is converted to a 2D representation by stacking the truncated spectra belonging to different channels to get a matrix of size $(B \cdot b)\times T$ , where b, T and B are the number of truncated bins, number of frames in the MS and number of filter banks used to obtain the MS representation, respectively. Thus for every frame, this representation has $(B \cdot b)$ dimensional features which are referred to as MS features. For evaluation, the Wiener filter is obtained using the procedure depicted in Fig. 1 with MS features as the input features. Since phase information in the MS is disregarded (non-negativity), signal reconstruction is not unique. For instance, any circular temporal shift (modulo the window length) of the DFT will lead to the same MS exemplar. However, this ambiguity can be reduced greatly if the magnitude spectrogram is sampled fast enough using smaller hop sizes. Even though using smaller hop sizes to obtain the MS features lead to temporal oversampling, it is found to be useful for making the mapping nearly one-to-one and make it useful for the proposed setup. This system is referred to as the Coupled MS system. To our knowledge, this is the first use of MS features for exemplar-based speech enhancement purpose.

## 4. RESULTS

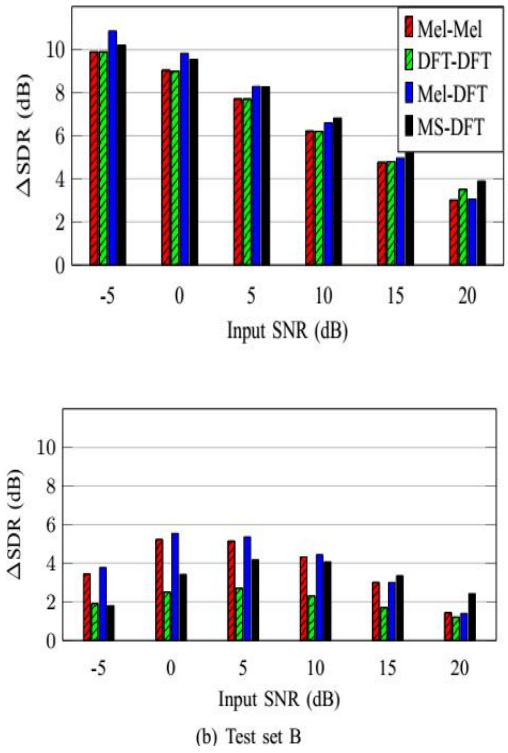


Figure 3: Average SDR improvements in dB obtained on test sets A and B of the AURORA-2 database as a function of input SNRs in dB for various settings. Legends are same for both plots

## 5. CONCLUSION

In this work, we presented a novel method to address the lowrank approximation of the estimates obtained in an exemplar-based speech enhancement system which uses features other than the fullscale DFT features. It has also been shown that the proposed system with coupled dictionaries can be made useful for features where a direct conversion from the feature space to the DFT space is not possible; for e.g. the MS features. The simulation results revealed that the proposed system yields better performance when compared to the baseline systems in terms of SDR. This is the first use of modulation envelope features for the exemplar-based speech enhancement purpose. The paper also presented a comparative study between the speech and noise separation capabilities of various feature representations.

## REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 27, no. 2, pp. 113–120, 1979.

[2] V. Grancharov, J. Samuelsson, and Bastiaan Kleijn, "On causal algorithms for speech enhancement," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 3, pp. 764–773, 2006.

[3] J.R. Jensen, J. Benesty, M.G. Christensen, and S.H. Jensen, "Enhancement of single-channel periodic signals in the timedomain," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 7, pp. 1948–1963, 2012.

[4] T.V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," Speech and Audio Processing, IEEE Transactions on, vol. 4, no. 5, pp. 383–389, 1996.

[5] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," Signal Processing, IEEE Transactions on, vol. 40, no. 4, pp. 725–735, 1992.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 3, pp. 1066–1074, 2007.

[7] G.J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech

from noise with the use of temporal dynamics," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, 2011, pp. 17–20.

[8] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplarbased sparse representations for noise robust automatic speech recognition," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 7, pp. 2067–2080, 2011.

[9] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 21, no. 10, pp. 2140–2151, 2013.

[10] T. Virtanen, R. Singh, and B. Raj, Eds., Techniques for Noise Robustness in Automatic Speech Recognition, Wiley, 2012.

**Mr. G. Lakshma Reddy** was born in Guntur, AP. He is graduated from the Jawaharlal Nehru Technological University, Hyderabad, His employment experience included PrakasamEngineeirngCollege, Kandukur, the Nalanda Institute of Engineering and technol-ogy, and Institute for Electronic Governance, Hyderabad. His special fields of interest included VLSI & Embedded Systems, Digital Signal Processing & communication Systems. Presently He is working as a Asst Prof in Newton's Institute of Engineering, Macherla. So far he is having 8 Years of Teaching Experience in various reputed engineering colleges.

Mr. KAKARLA ASHOK KUMAR was born in prakasam, AP on June 06, 1993 . He graduated from the Jawaharlal Nehru Technological University, Kakinada. . His special fields of interest included Communication Systems. Presently He is studying M.Tech in Newton's Institute of Engineering, Macherla.