INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A class-specific speech enhancement for phoneme recognition: a dictionary learning approach

Nazreen P.M.[1], *A. G. Ramakrishnan*[1], *Prasanta Kumar Ghosh* [2]

[1]Medical Intelligence and Language Engineering (MILE) Laboratory.
[2] Signal Processing Interpretation and Representation (SPIRE) Laboratory.
Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India.
{nazreenpm, ramkiag, prasantg} @ee.iisc.ernet.in

Abstract

We study the influence of using class-specific dictionaries for enhancement over class-independent dictionary in phoneme recognition of noisy speech. We hypothesize that, using class-specific dictionaries would remove the noise more compared to a class-independent dictionary, thereby resulting in better phoneme recognition. Experiments are performed with speech data from TIMIT corpus and noise samples from NOISEX-92 database. Using KSVD, four types of dictionaries have been learned: class-independent, manner-of-articulation-class, place-of-articulation-class and 39 phoneme-class. Initially, a set of labels are obtained by recognizing the speech, enhanced using a class-independent dictionary. Using these approximate labels, the corresponding class-specific dictionaries are used to enhance each frame of the original noisy speech, and this enhanced speech is then recognized. Compared to the results obtained using the class-independent dictionary, the 39 phoneme-class based dictionaries provide a relative phoneme recognition accuracy improvement of 5.5%, 3.7%, 2.4% and 2.2%, respectively for factory2, m109, leopard and babble noises, when averaged over 0, 5 and 10 dB SNRs.

Index Terms: speech enhancement, robust speech recognition, sparse coding, dictionary learning, phoneme recognition.

1. Introduction

In the past decade, there has been tremendous improvements in the field of automatic speech recognition (ASR). Despite these, the performance of an ASR system degrades significantly in the presence of noise due to the mismatch between the training and test environments, for example, when training is done on clean speech and testing is performed on noisy speech. The presence of noise distorts the spectrum of speech and hence degrades the performance.

Several techniques have been proposed to address this problem, and improve the recognition performance in noisy environments. One such method is to employ model adaptation schemes, like parallel model combination [1] and HMM adaptation [2, 3, 4]. Another approach is to analyze the existing features and enhance them to make them more noise robust, like cepstral mean subtraction [5], RASTA filtering [6] and vector Taylor series [7]. A third approach is to enhance the speech as a front end processing, using methods such as spectral subtraction [8] or Wiener filtering [9] before it is fed into a recognizer. This obviates the need to retrain the ASR systems for different types of noisy inputs since the same ASR trained on clean speech can be used. A comparative study [10] has also been reported on the performance of ASR system with various enhancement approaches. Recently, sparse coding techniques have gained popularity. A speech enhancement scheme based on sparse coding has been proposed by Sigg *et al.* [11], who show that it performs better than techniques like geometric spectral subtraction[12]. Several exemplar based techniques [13, 14] have also been proposed in the past for robust speech recognition.

In sparse coding, the basic assumption is that we can represent structured signals like speech as sparse linear combinations of prototype vectors or basis. Speech signal is composed of several sounds which can be categorized in various ways, like manner-of-articulation (MOA) [15], place-of-articulation (POA) [16, 17] or phonemes (PHN). Some of these classes might correlate well with certain noise types more than the other classes. Hence the bases in a dictionary learned using these classes may represent noise power to varying degrees and consequently result in poor speech reconstruction. By removing the contribution from bases of these classes that correlate well with noise, one could improve the enhancement performance. One way to achieve this is to learn different dictionaries for different classes and intelligently select a particular dictionary for a segment. Raj *et al.* [18] propose a similar approach, where they use phoneme-dependent non-negative matrix factorization (NMF) for separation of music from speech. In this work, we extend their idea to sparse coding to analyze how, using class-specific dictionaries, the performance of an ASR system could be improved over that obtained using a dictionary learned in a class-independent manner. Wang et al [19] investigated the use of class-specific, ideal ratio mask estimation for speech enhancement. But the recognizer used as well as the mask estimator are trained using noisy speech. However, we consider a more realistic scenario where the noise level is not known apriori and a recognizer trained on clean speech is used.

2. Enhancement using learned dictionary

Under additive model, noisy speech can be represented as,

$$y_t(m) = s_t(m) + n_t(m) \tag{1}$$

where $y_t(m)$, $s_t(m)$ and $n_t(m)$ are the m^{th} samples of the time domain noisy speech, clean speech and noise, respectively. Considering the short time Fourier transform (STFT),

$$y(\omega_k) = s(\omega_k) + n(\omega_k) \tag{2}$$

where $\omega_k = \frac{2\pi k}{R}$, $k = 0, 1, 2...R-1$, R is the number of frequency bins and k is the index. Taking the magnitude STFT, the noisy speech can be approximated as $y \approx s + n \in \mathbb{R}^{R\times 1}$, where s and n represent the spectra of the clean speech and the noise, respectively. An estimate of the STFT of the noisy speech is given by

$$\hat{y} = D_s \times c_s + D_n \times c_n \tag{3}$$

http://dx.doi.org/10.21437/Interspeech.2016-236

where $D_s \in \mathbb{R}^{R\times L}$ and $D_n \in \mathbb{R}^{R\times L}$, $L > R$, denote the speech and noise overcomplete dictionaries of L atoms each. c_s and c_n are the corresponding sparse coefficient vectors. Thus the enhanced speech is estimated as $\hat{s} = D_s \times c_s$.

2.1. Sparse coding

For a given dictionary D and the spectrum y of a given noisy speech frame, the sparse coefficients can be obtained by solving

$$c_o^* = \underset{c_o}{\operatorname{argmin}} \|y - Dc_o\|_2; \text{ s.t. } \|c_o\|_0 \leq t;\ t \ll R \quad (4)$$

The above problem can be solved by various schemes like orthogonal matching pursuit [20], which is a greedy iterative approach. Applying a convex relaxation of ℓ_0 norm to ℓ_1 norm, the problem becomes

$$c_o^* = \underset{c_o}{\operatorname{argmin}} \|y - Dc_o\|_2; \text{ s.t. } \|c_o\|_1 \leq t_1;\ t_1 \ll R \quad (5)$$

This formulation is known as least absolute shrinkage and selection operator (LASSO) [21]. Least angle regression (LARS) [22] is a very efficient algorithm, which gives a solution very close to LASSO. For the present work, we use LARS with a slight modification called batch LARS with coherence criterion (LARC) [11]. In LARC, a threshold is applied on the residual coherence as a stopping criterion.

2.2. Dictionary learning

We have used K-singular value decomposition (KSVD) based dictionary learning [23]. It is an iterative algorithm, which tries to sparsely represent a given data matrix X. It involves both sparse coding and dictionary update stages. The algorithm tries to solve the following problem,

$$\min_{D,C} \|X - DC\|_F^2 ; \text{ s.t. } \|c_i\|_0 \leq t\ \forall i;\ t \ll R \quad (6)$$

where $\|.\|_F^2$ indicates the squared Frobenius norm. We use an approximate KSVD [24], with reduced complexity.

3. Phoneme recognition on speech enhanced with class specific dictionaries

The performance of an ASR system on the enhanced speech depends not only on how much noise reduction is obtained but also on the amount of distortion of the speech components in the enhanced speech. We analyze how the ASR performance varies when we use class-specific dictionaries for enhancement rather than a class-independent one. Figure 1 shows the block diagram summarizing the steps of class-specific dictionary based enhancement for phoneme recognition proposed in the present study. At first, the class label of each frame is obtained by recognizing the speech enhanced using the class-independent dictionary. Using this approximate label, the corresponding class-specific dictionary, which was learned from the training data, is used to enhance the noisy speech in each frame, and this enhanced speech is recognized again. Three different categories of dictionaries are considered. Let there be c dictionaries in each category. In the first category, separate dictionaries are learned based on manner-of-articulation of speech where $c = 5$, denoted by $D_1^{MOA}, ...D_5^{MOA}$. In the second category, dictionaries are learned based on place-of-articulation of speech where $c = 14$, denoted by $D_1^{POA}, ...D_{14}^{POA}$. In the third case, separate dictionaries are learned for 39 different phonemes [25, 26] with $c = 39$, denoted by $D_1^{PHN}, ...D_{39}^{PHN}$. The enhancement and recognition stages are explained in Algorithm1.

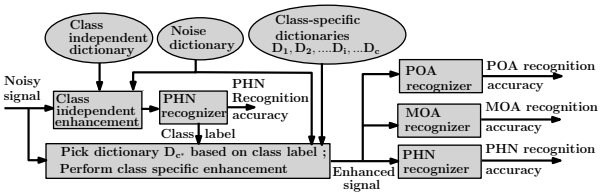

Figure 1: Phoneme recognition on speech enhanced with class specific dictionaries

Algorithm 1

1. *Enhance the noisy data using a class-independent dictionary:*
   Let $y \in \mathbb{R}^{R\times 1}$ be the noisy speech spectrum. $D_{ind} \in \mathbb{R}^{R\times L}$ and $D_n \in \mathbb{R}^{R\times L}$ be the dictionaries for class-independent speech and the noise, respectively. Using the composite dictionary $D = [D_{ind}\ D_n]$, the sparse coefficients of the noisy speech are obtained as
   $$[c_s^{ind}\ c_n] = LARC(y, D, \mu_{coh}) \quad (7)$$
   where μ_{coh} is the threshold on mutual coherence and c_s^{ind} represents the sparse coefficient vector corresponding to D_{ind}. Clean speech is estimated as
   $$\hat{s} = D_{ind} \times c_s^{ind} \quad (8)$$
2. *Find the phoneme labels using a phoneme recognizer on this enhanced speech. From the phoneme labels, obtain both the MOA and POA class labels of each frame.*
3. *Perform class-specific enhancement of the original noisy data using the dictionary corresponding to the obtained class label:*
   Three different enhancements are carried out based on the MOA, POA and PHN labels of the frame obtained from step 2.
   Method 1: *In this method, depending on the MOA class label the enhanced speech observation $\hat{s}$ is assigned to, the corresponding dictionary is chosen for enhancing the original noisy speech observation y. Let the class label be c^*; $1 \leq c^* \leq 5$. Thus, the sparse coefficients and the clean speech estimate obtained using composite dictionary $D1 = [D_{c^*}^{MOA}\ D_n]$ are*
   $$[c_s^{MOA}\ c_n^{MOA}] = LARC(y, D1, \mu_{coh}) \quad (9)$$
   $$\hat{s}^{MOA} = D_{c^*}^{MOA} \times c_s^{MOA} \quad (10)$$
   where c_s^{MOA} corresponds to $D_{c^*}^{MOA}$
   Method 2: *In this method, we use dictionaries based on POA, depending on the assigned label c^*; $1 \leq c^* \leq 14$, of $\hat{s}$. The sparse coefficients and the clean speech estimate obtained using the composite dictionary $D2 = [D_{c^*}^{POA}\ D_n]$ are*
   $$[c_s^{POA}\ c_n^{POA}] = LARC(y, D2, \mu_{coh}) \quad (11)$$
   $$\hat{s}^{POA} = D_{c^*}^{POA} \times c_s^{POA} \quad (12)$$
   where c_s^{POA} corresponds to $D_{c^*}^{POA}$
   Method 3: *This method employs dictionaries based on the assigned PHN labels; $1 \leq c^* \leq 39$, of $\hat{s}$. Using the composite dictionary $D3 = [D_{c^*}^{PHN}\ D_n]$, the sparse coefficients and the clean speech are estimated as*
   $$[c_s^{PHN}\ c_n^{PHN}] = LARC(y, D3, \mu_{coh}) \quad (13)$$
   $$\hat{s}^{PHN} = D_{c^*}^{PHN} \times c_s^{PHN} \quad (14)$$
   where c_s^{PHN} corresponds to $D_{c^*}^{PHN}$

4. *Find the performance of the MOA, POA and PHN recognizers on the enhanced speech in each case;* (10)*,* (12) *and* (14)*.*

4. Experiments and results

4.1. Experimental setup

All the experiments are conducted on the TIMIT [27] speech corpus consisting of 6300 sentences from 630 speakers with train and test sets containing 4620 and 1680 utterances, respectively. The sampling frequency is 16 kHz. The *sa* utterances are not used, since they are common to both training and testing sets. The ASR is trained on the entire clean TIMIT training data and its performance is obtained on the entire TIMIT test set. We use factory2, m109, leopard, babble and volvo noises from the NOISEX-92 [28] database after downsampling to 16 kHz, to synthesize noisy test speech signals. For the recognition experiments HTK [29] is used. The size of analysis frame is chosen to be 30 ms with 10 ms frame shift. 39-dimensional mel frequency cepstral coefficients (MFCC) [30] are used with zeroth coefficient, delta and delta-delta coefficients. Cepstral mean normalization (CMN) is applied. A three-state monophone HMM model with diagonal covariance matrix is used for the recognizer. The number of Gaussian mixtures per state is set to 32, since increasing it further does not improve the recognition performance significantly. A bigram phoneme language model is used. For phoneme recognizer, the 61 phonemes in TIMIT are mapped to a reduced set of 39 phonemes [25, 26] and the results are reported on this reduced set.

The dictionaries are learned on the magnitude STFT computed using a frame size of 30 ms with 10 ms frame shift. A 512-point FFT is taken and we use only the first 257 points for learning the dictionary because of symmetry in the spectrum. We use approximate KSVD algorithm with LARC coding [11] for learning the dictionaries. The number of iterations for KSVD is set to 30. The dictionaries are speaker independent and each dictionary is trained to have 512 basis vectors. The class-independent dictionary is learned on a subset of 2×10^5 frames, which are randomly sampled from the training data. For learning class-specific dictionaries, the training frames are classified into different classes, using the TIMIT labels. MOA, POA, as well as PHN specific dictionaries are learned from the spectra of corresponding training frames. For MOA class, vowels, diphthongs and semivowels are grouped together [15]. For POA class, consonants and vowels are classified as per [16] and [17], respectively. For PHN specific dictionary, we learn only 39 dictionaries based on the reduced phoneme set.

4.2. Results and discussion

Improvements in the phoneme recognition accuracies are compared across the three enhancement methods for 0, 5 and 10 dB SNRs. Figures 2 (a-e) show the phoneme recognition accuracies for factory2, m109, leopard, babble and volvo noises, respectively. We compare the recognition accuracies of our method with class-independent enhancement scheme and also with four other enhancement schemes; multi-band spectral subtraction (MBSS) [31], non causal apriori SNR estimator (NC) [32], harmonic regeneration noise reduction (HRNR) [33] and geometric spectral subtraction (GA) [12]. Our method achieves superior performance over all the methods.

Figure 2 shows that enhancement using class-specific dictionaries outperforms the class-independent enhancement in terms of phoneme recognition accuracies. This is true not only when we use class labels from the ground truth but also from the recognition of speech enhanced using class-independent dictionary (referred to as approximate labels). For phoneme recognition, PHN based enhancement using approximate labels gives a relative accuracy improvement (RAI) of 5.5%, 3.7%, 2.4% and 2.2%, respectively for factory2, m109, leopard and babble noise over class-independent enhancement method, when averaged over SNRs 0, 5 and 10 dB. MOA based enhancement gives average RAI of 2.7%, 2.3%, 1.6% and 2.2%, respectively. Similarly for POA based enhancement, the average RAIs are 4.3%, 2.5%, 1.8% and 2.1%.

The recognition accuracies obtained from the speech enhanced using ground truth labels (Figure 2), show that, we get higher performance as the number of classes c increases. It is to be noted that c_s in Eq. (3) need not be zero even if the speech component in y is zero. We refer this contribution of speech bases for representing noise as noise confusion. We observe that, as we increase the number of classes and use only one class dictionary per frame the noise confusion reduces[1]. When we use approximate labels, the performance improvement also depends on the accuracy of ASR, which usually goes down as the number of classes increases. Hence to achieve the best recognition performance, one needs to choose an optimal number of classes by trading off ASR accuracy and noise confusion. It is observed that, PHN based enhancement outperforms MOA and POA based enhancements in most cases. This indicates that the approximate PHN labels obtained from the ASR are good enough to get a performance better than that from the MOA and POA labels.

For babble noise, at 0 dB SNR, no significant improvement is observed when we use the approximate labels. This could be due to the very low recognition accuracy that we obtain after the enhancement using class-independent dictionary resulting in a poor choice of dictionary for most frames.

In the case of volvo noise, it is observed that after CMN, the recognition accuracy using noisy speech outperforms the class-independent and class-dependent schemes. For phoneme recognition, our PHN based enhancement using approximate labels shows an average relative degradation of -0.8% over the noisy performance. However, it is to be noted that the results for class-dependent schemes are still better than the class independent scheme. For phoneme recognition, the average RAIs over class-independent scheme are 2.2%, 1.6% and 1.6% for PHN, MOA and POA based enhancements, respectively. However when we use the PHN labels obtained from the noisy speech itself, PHN based enhancement using approximate labels gives an average RAI of 2.1% over the noisy case.

Figure 3 shows the log magnitude spectral plots of a few exemplary frames which are correctly recognized after PHN based enhancement using ground truth labels (PHN-gnd) but wrongly recognized after class independent enhancement, for the noises factory2, m109 and leopard at 0 dB SNR. The spectral plots and the corresponding Itakura-Saito (IS) distortion measures with the clean spectrum show that the spectrum recovered after PHN-gnd enhancement matches more closely with the clean speech spectrum than that after the class-independent enhancement.

Instead of performing enhancement followed by recognition only once, one can think of a multi-stage enhancement-

[1] A small experiment with a total of 300 Factory2 noise frames demonstrates that the fraction corresponding to the energy of the coefficients for class-independent dictionary is 0.025 when both class-independent and noise dictionaries are used for sparsely representing the noise frames. However, the fraction reduces to 0.0184 (averaged over all 39 phonemes classes) when the phoneme-specific dictionaries are used in place of a class-independent dictionary.

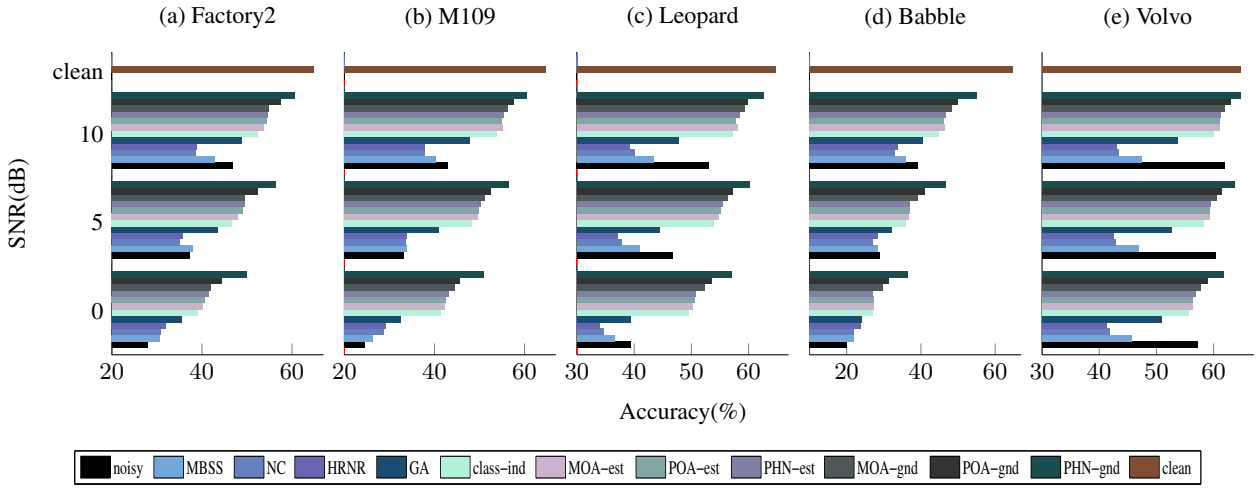

Figure 2: Comparison of phoneme recognition accuracies for (a) Factory2, (b)M109, (c) Leopard (d) Babble and (e) Volvo noises. In each figure, results are given for SNRs of 0, 5 and 10 dB. For each SNR, the recognition accuracies are given for noisy speech, speech enhanced using MBSS, NC, HRNR, GA, class independent enhancement scheme and various class dependent enhancement schemes. In the legend, class-ind, MOA-est, POA-est, PHN-est, MOA-gnd, POA-gnd and PHN-gnd refer to class-independent case, MOA, POA, PHN enhancement using estimated labels , MOA, POA and PHN enhancement using ground truth labels, respectively

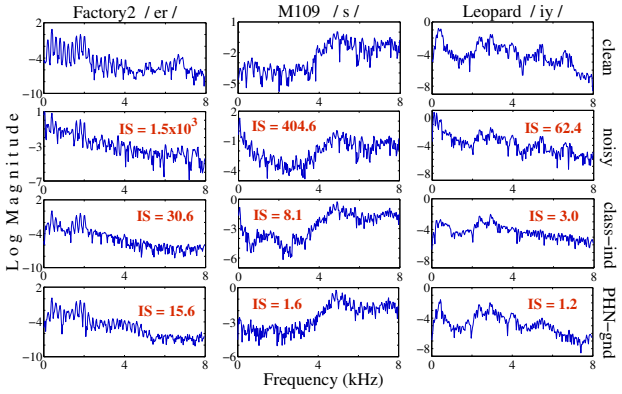

Figure 3: Log magnitude spectra of a few exemplary frames correctly recognized after PHN based enhancement using ground truth labels (PHN-gnd) but misclassified after class-independent (class-ind) enhancement, for different noises at 0 dB SNR. Rows from top to bottom correspond to the spectra of clean, noisy, class-ind and PHN-gnd speech, respectively. The Itakura-Saito (IS) distance of each spectrum from the clean spectrum is also shown. The phoneme and noise types of each exemplary frame is given at the top of the column.

recognition scheme, where class-specific enhancement is performed in each stage and the required class labels are taken from the recognition output of the previous stage. Experiment with a two-stage scheme, with PHN enhancement demonstrates that the second stage gives a relative improvement in phoneme recognition accuracies over the first stage by 1.0%, 0.8%, 1.0% and 0.7% (averaged over 0, 5, and 10 dB SNRs) for factory2, m109, leopard and babble noises, respectively.

We also analyze the performance improvements for MOA and POA recognizers for the single-stage scheme. In the case of MOA recognition, PHN based enhancement gives average relative recognition accuracy improvements of 2.4%, 2.1%, 2.3% and 2.8% for factory2, m109, leopard and babble noises, respectively, over class-independent enhancement, while MOA based enhancement gives improvements of 1.3%, 1.1%, 1.7% and 2.1%. Also for POA based enhancement, the average RAIs are 1.7%, 0.8%, 1.3% and 2.4%, respectively.

For POA recognition, PHN based enhancement gives average RAIs of 2.9%, 2.1%, 2.4% and 0.7%. MOA based enhancement achieves improvements of 2.3%, 1.8%, 1.9% and 0.9%. For POA based enhancement, we get improvements of 3.3%, 2.2%, 2.7% and 0.8%. This suggests that except for POA recognition, the PHN based enhancement yields better recognition accuracy than MOA and POA based enhancements.

5. Conclusions

We have analyzed how the recognition performance of noisy speech varies when we use class-specific dictionaries for enhancement rather than a class-independent dictionary. The experiments are carried out in a speaker independent scenario. With the ground truth class labels, there is significant improvement in recognition accuracy for class-specific enhancement over the class-independent scheme. When ground truth labels are used, the 39-PHN based enhancement gives average RAI in phoneme recognition of 21.5%, 17.6%, 12.1%, 29.2% and 9.3% for factory2, m109, leopard, babble and volvo noises, respectively, over class-independent enhancement.

The 39-PHN based enhancement outperforms the MOA and POA based schemes in most of the cases. Using the approximate labels obtained from the ASR gives better recognition accuracy than the class-independent enhancement, although it is lower than that using the ground truth labels. In our future work, we intend to employ an MLP-HMM framework with other features like multistream features [34] for recognition and examine the benefit of class-specific enhancement, since it has been shown to perform significantly better than GMM-HMM framework with MFCC features. Also we would like to examine the usefulness of our algorithms on a more realistic scenario involving real world speech mixed with noise.

6. References

[1] M. J. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *Speech and Audio Processing, IEEE Trans.*, vol. 4, no. 5, pp. 352–359, 1996.

[2] D. Pei and C. Zhigang, "An efficient robust automatic speech recognition system based on the combination of speech enhance-

ment and log-add HMM adaptation," in *Info-tech and Info-net. Proceedings. ICII 2001-Beijing. Int. Conf.*, vol. 3. IEEE, 2001, pp. 367–371.

[3] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *Speech and Audio Processing, IEEE Trans.*, vol. 13, no. 3, pp. 412–421, 2005.

[4] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions," *Computer Speech & Language*, vol. 23, no. 3, pp. 389–405, 2009.

[5] S. Furui, "Cepstral analysis technique for automatic speaker verification," *Acoustics, Speech and Signal Processing, IEEE Trans.*, vol. 29, no. 2, pp. 254–272, 1981.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *Speech and Audio Processing, IEEE Trans.*, vol. 2, no. 4, pp. 578–589, 1994.

[7] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing. ICASSP. Proceedings. IEEE Int. Conf.*, vol. 2, 1996, pp. 733–736.

[8] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech Signal Proc, IEEE Trans.*, vol. 27, no. 2, pp. 113–120, 1979.

[9] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Acoustics, Speech, and Signal Processing. ICASSP. Proceedings. IEEE Int. Conf.*, vol. 3, 2000, pp. 1875–1878.

[10] K. K. Paliwal, J. G. Lyons, S. So, A. P. Stark, and K. K. Wójcicki, "Comparative evaluation of speech enhancement methods for robust automatic speech recognition," in *4th Int. Conf. on Signal Processing and Communication Systems*, 2010.

[11] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *Audio, Speech, and Language Processing, IEEE Trans.*, vol. 20, no. 6, pp. 1698–1712, 2012.

[12] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech communication*, vol. 50, no. 6, pp. 453–466, 2008.

[13] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Trans.*, vol. 19, no. 7, pp. 2067–2080, 2011.

[14] E. Yilmaz, J. F. Gemmeke *et al.*, "Noise-robust speech recognition with exemplar-based sparse representations using alpha-beta divergence," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE Int. Conf.*, 2014, pp. 5502–5506.

[15] P. Scanlon, D. P. Ellis, and R. B. Reilly, "Using broad phonetic group experts for improved speech recognition," *Audio, Speech, and Language Processing, IEEE Trans.*, vol. 15, no. 3, pp. 803–812, 2007.

[16] "The International Phonetic Alphabet," revised to 2005.

[17] R. Rasipuram *et al.*, "Multitask learning to improve articulatory feature estimation and phoneme recognition," Idiap, Tech. Rep., 2011.

[18] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent NMF for speech enhancement in monaural mixtures." in *INTERSPEECH*, 2011, pp. 1217–1220.

[19] Z.-Q. Wang, Y. Zhao, and D. Wang, "Phoneme-specific speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[20] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers. Record of The Twenty-Seventh Asilomar Conf.* IEEE, 1993, pp. 40–44.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[22] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[23] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Trans.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[24] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," *CS Technion*, vol. 40, no. 8, pp. 1–15, 2008.

[25] P. MomayyezSiahkal, "Integration of multiple feature sets for reducing ambiguity in automatic speech recognition," Ph.D. dissertation, McGill University, 2008.

[26] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *Acoustics, Speech and Signal Processing, IEEE Trans.*, vol. 37, no. 11, pp. 1641–1648, 1989.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, Feb. 1993.

[28] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition ii: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

[29] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[30] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Trans.*, vol. 28, no. 4, pp. 357–366, 1980.

[31] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE international conference on acoustics speech and signal processing*, vol. 4. Citeseer, 2002, pp. 4164–4164.

[32] I. Cohen, "Speech enhancement using a noncausal a priori SNR estimator," *Signal Processing Letters, IEEE*, vol. 11, no. 9, pp. 725–728, 2004.

[33] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2098–2108, 2006.

[34] S. K. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Trans.*, vol. 21, no. 2, pp. 416–426, 2013.