A Joint Enhancement-Decoding Formulation for Noise Robust Phoneme Recognition

Nazreen P. M.
Medical Intelligence and Language Engineering (MILE) Laboratory.
Electrical Engineering
Indian Institute of Science (IISc)
Bangalore, India.
nazreenpm@ee.iisc.ernet.in

A. G. Ramakrishnan
Medical Intelligence and Language Engineering (MILE) Laboratory.
Electrical Engineering
Indian Institute of Science (IISc)
Bangalore, India.
ramkiag@ee.iisc.ernet.in

Prasanta Kumar Ghosh
Signal Processing Interpretation and Representation (SPIRE) Laboratory.
Electrical Engineering
Indian Institute of Science (IISc)
Bangalore, India.
prasantg@ee.iisc.ernet.in

***Abstract*—We consider a dictionary based speech enhancement in the context of automatic recognition of noisy speech. Speech in each analysis frame is denoised as a front-end processing using a class-specific (e.g., phoneme) dictionary selected based on the estimated class label. However, when the estimated label is erroneous, a wrong class model is chosen for many frames. We propose a Joint Enhancement-Decoding (JED) algorithm to overcome this issue by jointly optimizing for labels of all the frames and the decoding path. The algorithm optimizes over multiple enhanced versions of each frame using different phoneme specific dictionaries and gives the maximum likelihood path of state sequences as well as the best (in the maximum likelihood sense) choice of the enhanced observation sequence as its output. The number of phoneme-specific dictionaries used for enhancement in an analysis frame is varied from 1 to 5 based on the phoneme confusion matrix and the recognition results are reported for each case. Experiments with TIMIT corpus and five different noises at 0, 5 and 10 dB SNRs show that the recognition performance varies with the number of dictionaries, and in most of the cases, is the best when two or three dictionaries are employed.**

***Index Terms*—speech enhancement, robust speech recognition, sparse coding, dictionary learning.**

I. Introduction

Despite the improvement in the performance of automatic speech recognition (ASR) systems over the last few decades, the accuracy obtained over noisy test conditions is still poor. The presence of noise in the test data distorts the spectrum, thereby reducing the recognition performance.

The earliest techniques proposed to address this problem and improve the performance in noisy environments include parallel model combination [1], HMM adaptation [2][3][4], cepstral mean subtraction [5] and vector Taylor series [6]. Another approach is to enhance the speech as a front end processing before it is fed into the recognizer, thereby obviating the need to retrain the ASR system for different noise statistics [7] [8]. A comparative study on the performance of ASR system for various enhancement schemes has been reported in [9]. Sigg *et al.* [10] proposed a speech enhancement scheme based on sparse coding and showed that it performs better than techniques like geometric spectral subtraction [11]. Several exemplar based techniques [12][13] have also been proposed for noise-robust speech recognition.

Speech signal is composed of several sound classes which can be categorized into phonemes (PHN). A given noise type may correlate more with a few of these classes than the rest. The bases in a dictionary learned using these classes may represent noise power to varying degrees. Hence these bases may leak noise power into the enhanced speech and cause poor speech reconstruction [14]. By removing the contribution from bases of these classes that correlate well with noise, one could improve the enhancement performance. Raj *et al.* [15] proposed an approach using this concept, where they use phoneme-dependent non-negative matrix factorization (NMF) for separation of music from speech. Nazreen *et al.* [14] extended this idea to sparse coding by using class-specific dictionaries and compared its performance to that of a class-independent dictionary. Wang et al [16] investigated the use of class-specific, ideal ratio mask estimation for speech enhancement. But the recognizer as well as the mask estimator were trained using noisy speech.

All the class specific enhancement schemes mentioned so far depend on the estimated class label for each frame, which may be erroneous. This leads to the selection of wrong class model for the enhancement in the respective frames. To overcome this, we propose a Joint Enhancement-Decoding (JED) algorithm that jointly optimizes these class labels and the final recognized speech labels. By this approach, we aim to find the best possible frame-wise model for enhancement and the recognition labels together for an input speech signal in a single optimization framework. We develop this algorithm by integrating the class label estimation into the Viterbi decoder [17] typically used for speech recognition. We implement the same using the HTK toolkit. The proposed algorithm accepts multiple enhanced observations and chooses the best in each frame such that the overall likelihood is maximized. Multiple observations are obtained by enhancing every noisy frame using multiple class-specific dictionaries. The best sequence of observations is chosen to maximize the likelihood. Thus we don't separately choose a class label and consequently the

class model for frame-wise enhancement as in [14] [15]or [16].

We analyze the performance of our algorithm on TIMIT database. We use the confusion matrix obtained from the recognition output of clean speech to select the pool of dictionaries. This results in an improvement in the performance in most of the cases compared to the enhancement using a class-independent dictionary. It is to be noted that when the number of dictionaries is set to 1, the algorithm becomes the same as the one in [14].

II. Joint Enhancement-Decoding algorithm for class-specific enhancement

Block 1 in Figure 1 shows a generic class-specific enhancement framework [14] [15]. The class label is estimated for each frame of the noisy speech, and the corresponding class-specific dictionary is used for enhancement. When this estimate of class label is erroneous, it selects the wrong class dictionary resulting in poor enhancement of the respective frames. We propose a joint enhancement-decoding (JED) formulation to compensate for this error. The block diagram of class-specific enhancement using the proposed JED algorithm is shown in block 2 of Figure 1. The algorithm accepts multiple enhanced observations in each frame and selects the best observation in each frame as well as the best state sequence that maximize the likelihood of the chosen observations. We use multiple class-specific dictionaries for enhancing a single frame and these different denoised versions of a frame are fed into the JED algorithm. We use sparse coding based dictionary learning approach for obtaining the enhanced speech observations. This approach involves learning of speech and noise dictionaries as well as a sparse coding stage for learning the coefficients.

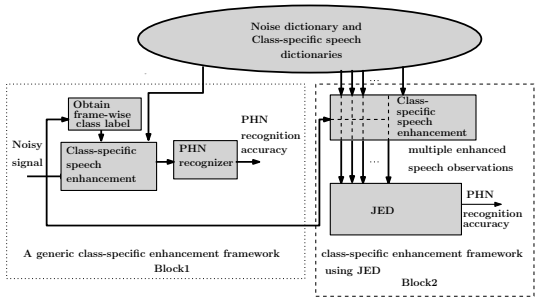

Fig. 1: Generic and JED based class-specific enhancement framework

A. Speech enhancement using learned dictionary

Let $y_t(m)$, $s_t(m)$ and $n_t(m)$ be the m^{th} samples of the noisy speech, clean speech and noise, respectively. Considering additive model, noisy speech can be represented as, $y_t(m) = s_t(m) + n_t(m)$. By taking the short time Fourier transform (STFT) we get, $y(\omega_k) = s(\omega_k) + n(\omega_k)$, 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. Considering only the magnitude spectra, we can write, $y \approx s + n \in \mathbb{R}^{R\times 1}$, where s and n represent the spectra of the clean speech and the noise, respectively. Using speech and noise dictionaries and their corresponding sparse coefficients, an estimate of the STFT of the noisy speech is given by $\hat{y} = D_s \times c_s + D_n \times c_n$ where D_s 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 $\hat{s} = D_s \times c_s$. For the present work, we use K-singular value decomposition (KSVD) based dictionary learning [18]. For sparse coding, we use batch LARS with coherence criterion (LARC) [10]. In LARC, a threshold is applied on the residual coherence as a stopping criterion.

B. JED Algorithm

Let the enhanced observation at the t^{th} frame using class-specific dictionary with i^{th} label be denoted by θ_t^i. If T denotes the total number of frames and N denotes the number of best labels considered for enhancement in each frame, $\Theta = \{\theta_t^i; 1 \le t \le T, 1 \le i \le N\}$. The JED algorithm optimizes the observation sequence $\theta_t^{i^*(t)}, 1 \le t \le T$ as well as the state sequence $s_1^*, s_2^*, \ldots, s_T^*$ to maximize the likelihood of the observation as follows :

$$\begin{aligned}\{\theta_t^{i^*(t)}, s_t^*, 1 \le t \le T\} &= \underset{\theta_t^i, s_t}{\operatorname{argmax}} P(s_1, s_2, \ldots s_T|\Theta) \\ &= \underset{\theta_t^i, s_t}{\operatorname{argmax}} P(\Theta|s_1, s_2, \ldots s_T)P(s_1, s_2, \ldots s_T) \\ &= \underset{s_t}{\operatorname{argmax}} \Big\{ \max_{\theta_t^i} P(\Theta|s_1, s_2, \ldots s_T) \Big\} P(s_1, s_2, \ldots s_T)\end{aligned} \tag{1}$$

Assuming independence among observations, given the state sequence, we write

$$\begin{aligned}\{s_1^*, s_2^*, \ldots s_T^*\} &= \underset{s_t}{\operatorname{argmax}} \Big\{ \prod_{t=1}^{T} \max_{1\le i\le N} P(\theta_t^i|s_t) \Big\} P(s_1, s_2, \ldots s_T) \\ &= \underset{s_t}{\operatorname{argmax}} \Big\{ \prod_{t=1}^{T} P(\theta_t^{i^*(t)}|s_t) \Big\} P(s_1, s_2, \ldots s_T)\end{aligned} \tag{2}$$

where $i^*(t) = \operatorname{argmax}_{1\le i\le N} P(\theta_t^i|s_t)$.

Algorithm 1 below gives the steps of the JED algorithm.

C. Best-N class-specific dictionary based enhancement using JED

JED algorithm does not require the knowledge of frame labels to do the class-specific enhancement and decoding. In fact, all the phoneme dictionaries can be used for enhancement of each frame and these multiple enhanced observations can be given as the input. The algorithm then jointly optimizes the class label in each frame and the decoding path. However the use of all phoneme dictionaries for frame-wise enhancement is computationally expensive. Hence we intelligently choose a subset of labels such that the chance of actual label belonging to this subset is high.

To choose this subset of labels, we use the confusion matrix obtained by running the recognizer on a subset of the clean TIMIT test sentences. For a given recognized label, selecting a small set of labels with high likelihood from the matrix

$\Theta = \{\theta_t^i; 1 \le t \le T, 1 \le i \le N\}$: observation sequence
$b(\cdot|s_j)$:observation probability given state s_j
$a(s_k \rightarrow s_j)$:transition probability from s_k to s_j

for *each state* $s! =$*Starting state* **do**
  $D(1, s) = 0$
end
for $t \leftarrow$ *1 to* T **do**
  for *each state* s_j **do**
    $\theta_t^{i^*(t)} = \text{argmax}_{\theta_t^i}\, b(\theta_t^i|s_j)$
  end
  for *each state* s_j **do**
    $D(t, s_j) \leftarrow \max_k D(t-1, s_k) b(\theta_t^{i^*(t)}|s_j) a(s_k \rightarrow s_j)$
    $\Psi(t, s_j) =$ $\text{argmax}_k\, D(t-1, s_k) b(\theta_t^{i^*(t)}|s_j) a(s_k \rightarrow s_j)$
  end
end
$P(\Theta, \mathbf{S}) = \max_j D(T, s_j)$
Backtrack

Algorithm 1: Joint Enhancement-Decoding Algorithm

ensures that the likelihood of the actual label being in this set is quite high. This assumption can be empirically seen to be true from Table I. Hence we use this set of labels in class specific enhancement for enhancing each frame in a noisy test speech.

Figure 2 shows the block diagram summarizing the steps of the best-N class-specific dictionary based enhancement for phoneme recognition using the proposed JED algorithm. At first, the phoneme label of each frame is estimated by recognizing the speech enhanced using a class-independent dictionary. The confusion matrix of ASR output for clean speech is used to obtain the next $N-1$ best labels. These labels are then used to obtain N enhanced observations for each frame using the respective dictionaries. These observations are then fed into the JED algorithm which gives the decoded output by maximizing the likelihood.

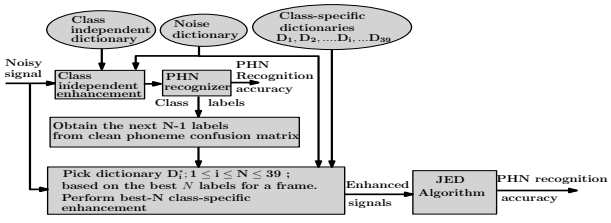

Fig. 2: Phoneme recognition of noisy speech using best-N class-specific dictionaries using JED

The enhancement and recognition stages are similar to that as Algorithm 1 in [14]. After obtaining the phoneme labels using a phoneme recognizer on a class-independent enhanced speech as in [14], we perform our Best-N class-specific dictionary based enhancement. The steps are as follows:

1) Based on the class label of a frame, obtain the next best $N-1$ labels using the phoneme confusion matrix obtained on a small subset of clean speech data.
2) Let the N best dictionaries corresponding to the obtained class label be D_i^*; $1 \le i \le N$. Enhance the original noisy speech observation y separately using each of these N best dictionaries. Thus, the sparse coefficients and the clean speech estimates obtained using the composite dictionary $D_i = [D_i^*\ D_n]$ are

$$\begin{bmatrix} c_s^{*i} \\ c_n^{*i} \end{bmatrix} = LARC(y, D_i, \mu_{coh}) \quad (3)$$

$$\hat{s^*}_i = D_i^* \times c_s^{*i} \quad (4)$$

3) Input these N enhanced estimates for each frame to the JED algorithm and evaluate the recognition performance.

III. Experiments and results

A. Experimental setup

For all the experiments we use TIMIT [19] 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. We use factory2, m109, leopard, babble and volvo noises from the NOISEX-92 [20] database after downsampling to 16 kHz, to synthesize noisy test speech signals at 0, 5 and 10 dB SNRs.

1) Enhancement setup: 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 only the first 257 points is used for learning the dictionary because of symmetry in the spectrum. The number of iterations for KSVD is set to 30. The dictionaries are speaker independent and contain 512 basis vectors each. The class-independent dictionary is learned on a subset of 2×10^5 frames, randomly sampled from the training data. The training frames are classified into different phoneme classes, using the TIMIT labels. Phoneme specific dictionaries are learned from the spectra of these frames. The 61 phonemes in TIMIT are mapped to a reduced set of 39 phonemes [21], [22]. We learn phoneme-specific dictionaries based on this reduced phoneme set.

2) Recognition setup: The ASR is trained on the entire clean TIMIT training data. The TIMIT test set is randomly divided into two equal sets. One of them is used to obtain the clean speech confusion matrix after recognition. The recognition accuracies are compared on the second test set. To implement our JED algorithm, we modified the source code of Viterbi decoding for recognition in the HTK toolkit [23]. The analysis frame is chosen to be 30 ms with 10 ms frame shift. 39-dimensional mel frequency cepstral coefficients [24] are used for recognition with 0-th coefficient, delta and delta-delta coefficients. Cepstral mean normalization is applied. A three-state monophone HMM model with diagonal covariance matrix is used for recognition. 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. The results are reported on the reduced phoneme set.

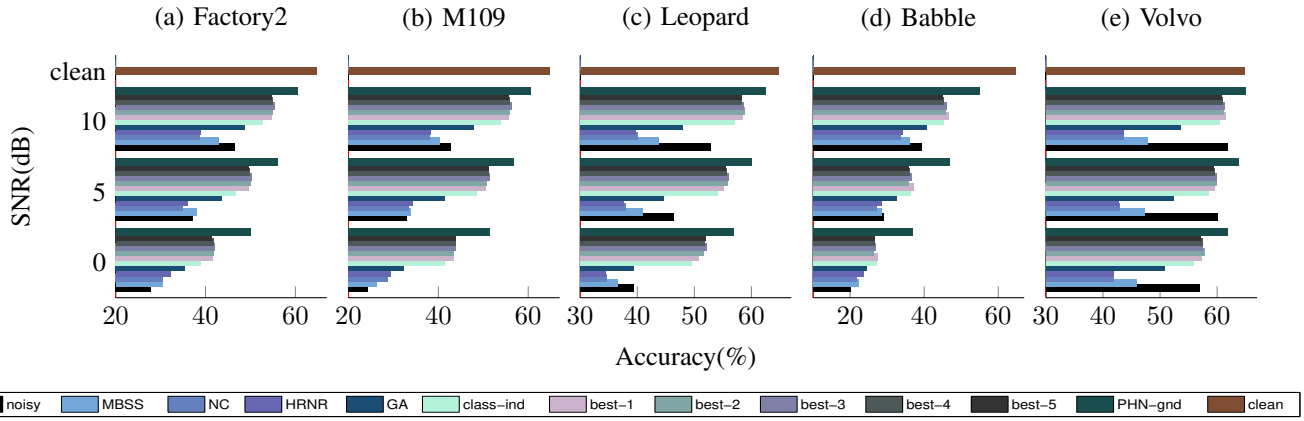

Fig. 3: Performance of JED in terms of phoneme recognition accuracies on speech with five different noises. In each figure, results are given for SNRs of 0, 5 and 10 dB. For each SNR, the recognition accuracies are given for noisy speech, speech enhanced using MBSS, NC, HRNR, GA, class independent enhancement scheme and best-N class dependent enhancement (best-N) schemes for N varying from 1 to 5. PHN-gnd refers to the ideal case, when the ground truth phoneme labels are used for enhancement.

B. Results and discussion

Improvements in the phoneme recognition accuracies are compared for the proposed best-N class-dependent enhancement scheme using JED for N varying from 1 through 5. Figure 3 shows the phoneme recognition accuracies for the five noises. We compare the recognition accuracies of the proposed method with class-independent enhancement scheme and also with four other enhancement schemes: multi-band spectral subtraction (MBSS) [25], non-causal apriori SNR estimator (NC) [26], harmonic regeneration noise reduction (HRNR) [27] and geometric spectral subtraction (GA) [11].

From figure 3, we notice that the best-N enhancement scheme yields accuracies superior to the class independent case for all noise types. We get a marked improvement over class-independent scheme when $N = 1$. However, the improvement for $N > 1$ is minimal over $N = 1$ case.

For factory2 noise, best-N enhancement scheme gives an average relative accuracy improvement (RAI) of 5.6%, 6.2%, 6.9%, 6.0% and 5.4% respectively, for values of $N = 1$ to 5, over class-independent enhancement scheme, when averaged over SNRs 0, 5, and 10 dB. For leopard noise, the average RAI values are 2.3%, 3.6%, 3.9%, 3.2% and 3.2%. The RAI values for M109 noise are 3.9%, 4.3%, 5.4%, 4.9% and 4.8% respectively.

In the case of babble noise, the proposed scheme gives superior performance only when $N = 1$. The average RAI values over class-independent scheme are 2.3%, -1.0%, 0.8%, -1.1% and -1.3% for values of $N = 1$ to 5.

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 in most cases. Thus the proposed scheme shows average RAI values of -0.2%, 0.1%, -0.02%, -0.4% and -0.8% respectively, for N varying from 1 to 5, over the noisy performance. However, it is to be noted that the accuracies from the proposed scheme are still better than those of the class independent scheme. For phoneme recognition, the average RAIs over class-independent scheme are 2.0%, 2.3%, 2.2%, 1.8% and 1.4% respectively, for N varying from 1 to 5.

TABLE I: Percentage of frames for which none of the estimated N labels are correct. The two columns for each noise correspond to N=1 and N=5.

SNR	Factory2		M109		Leopard		Babble		Volvo	
(dB)	1	5	1	5	1	5	1	5	1	5
0	52	30	49	28	42	24	69	43	37	20
5	45	25	43	23	39	21	57	33	35	19
10	40	21	38	21	37	20	47	27	34	18

Figure 3 shows that as N varies from 1 to 5, the recognition performance varies, giving the best with two or three dictionaries in most of the cases. The benefit of using multiple enhanced observations based on best-N class-specific dictionaries could be explained from the fact that the class labels employed for a frame have more chance of having the correct label when N=5 than when N=1. To illustrate this, we report the percentage of frames where the estimated labels do not include the ground truth class label for both N=1 and N=5 for different noise and SNR conditions in Table I. It is clear that the percentage of such frames reduces when N=5 compared to when N=1.

TABLE II: Log likelihood values of a few utterances for best-N class-dependent schemes (best-N) for N varying from 1 to 5 for factory2 noise at 0 dB SNR.

	mnjm0/sx410	fpas0/sx404	mtaa0/sx115	fcal1/sx143
best-1	-21820	-21494	-23795	-19125
best-2	-20853	-20885	-22755	-18644
best-3	-20299	-20380	-22139	-18187
best-4	-20078	-20040	-21985	-18121
best-5	-19909	-19781	-21876	-18074

TABLE III: Phoneme recognition accuracies for best-N ; $N =$ 2 to 5 using n-gram confusion matrix ($N = 1$ to 3) averaged over Factory 2, Babble, Leopard, M109 and Volvo noises for 0, 5 and 10 dB SNRs

	best-2	best-3	best-4	best-5
Single	50.0	50.3	49.9	49.8
Bigram	50.2	50.3	50.2	50.2
Trigram	50.2	50.4	50.3	50.2

As described in section II-C, the JED maximizes the overall likelihood of the output utterance. Table II shows the log likelihood values of a few utterances for best-N class-dependent schemes for factory2 noise at 0 dB SNR, from which we notice that the likelihood increases monotonically from N = 1 to 5. However, this does not always translate to monotonic increase in recognition accuracy as is evident from Fig 3.

To explore the effect of dependency of phonemes, we also repeated the experiments using a bigram and trigram confusion matrix. In the bigram case, the best-N phonemes for each frame was selected based on a bigram confusion matrix. This matrix was populated by computing the occurrence of each phoneme for a given combination of estimated phonemes at the current and previous time instant. Similarly, trigram confusion matrix uses the current, previous and next frames for computing the frequency of occurrence of a phoneme. Table III shows the recognition accuracies for this experiment when averaged over all the five noises and the three SNRs. It is observed that compared to the case of using a single confusion matrix, the bigram and trigram cases give only marginal improvements.

IV. Conclusions and Future work

We analyzed the phoneme recognition performance of JED using best-N class-specific dictionaries. The recognition performance varies with N, giving the best values at N = 2 or 3 in most cases. Further, the performance also depends on the type of noise corrupting the speech. Thus, in a real life scenario, the performance can be optimized by first identifying the label of the noise and then employing the relevant noise dictionary.

The input observations for JED algorithm need not necessarily be enhanced using class-specific dictionary based approaches. The recognition performance of different enhancement techniques varies substantially over different noise types and SNRs [9]. Hence one can choose any other denoising technique depending upon the noise type and SNR. The proposed algorithm can thus be used to find the best enhancement scheme and recognition label for an input speech with any noise. We intend to explore in this direction in future.

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 enhancement and log-add HMM adaptation," in *Info-tech and Info-net. Proc. ICII 2001-Beijing. Int. Conf.*, vol. 3. IEEE. IEEE, 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] 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. IEEE, 1996, pp. 733–736.

[7] 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.

[8] 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. IEEE, 2000, pp. 1875–1878.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.* IEEE, 2014, pp. 5502–5506.

[14] P. M. Nazreen, A. G. Ramakrishnan, and P. K. Ghosh, "A class-specific speech enhancement for phoneme recognition: A dictionary learning approach," *Proc. Interspeech*, pp. 3728–3732, 2016.

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

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

[17] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[18] 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.

[19] 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.

[20] 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. [Online]. Available: http://dx.doi.org/10.1016/0167-6393(93)90095-3

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

[22] 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.

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

[24] 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.

[25] 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.

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

[27] 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.