FEATURE ENHANCEMENT USING SPARSE REFERENCE AND ESTIMATED SOFT-MASK EXEMPLAR-PAIRS FOR NOISY SPEECH RECOGNITION

Lee Ngee Tan *Abeer Alwan*

Department of Electrical Engineering, University of California, Los Angeles, USA
{tleengee, alwan}@ee.ucla.edu

ABSTRACT

A feature enhancement technique for noise-robust speech recognition is proposed. Existing sparse exemplar-based feature enhancement methods use clean speech and pure noise Mel-spectral exemplars, or clean and noisy speech log-Mel-spectral exemplar-pairs, in their dictionaries. In contrast, the proposed technique constructs its dictionaries using reference soft-mask (SM_{ref}) and estimated soft-mask (SM_{est}) exemplar-pairs derived from the training data. The sparse linear combination of SM_{est} dictionary exemplars that best represents the test utterance's SM_{est} is obtained by solving an L1-minimization problem. This sparse linear combination is applied to the SM_{ref} exemplar dictionary to generate an enhanced soft-mask for denoising the utterance's Mel-spectra before MFCC extraction. On the Aurora-2 noisy speech recognition task, the proposed algorithm outperforms other sparse Mel-spectral exemplar-based feature enhancement schemes when mismatch exists between the dictionary exemplars and the test set. A preliminary experiment on Aurora-4 shows similar trends.

***Index Terms*—** Feature enhancement, soft mask estimation, noisy speech recognition, sparse exemplar, joint dictionary.

1. INTRODUCTION

Although good speech recognition performance has been achieved for clean speech, automatic speech recognition (ASR) in noise is still a challenging problem. Both front-end and back-end processing techniques have been proposed to tackle this problem. Front-end processing involves the extraction of noise-robust speech representations or features which are more invariant in noise. This can be done using speech or feature enhancement / denoising techniques (e.g. spectral subtraction [1] and soft-mask estimation [2, 3]), or extracting speech salient features (e.g. multi-scale spectro-temporal features [4] and normalized modulation cepstral coefficients (NMCC) [5]). Back-end processing involves model adaptation to reduce the mismatch between the trained models and the test conditions [6, 7]. With robust performance reported using compressive sensing / sparsity-based techniques in image processing applications [8, 9], the use of sparse spectral representations for feature enhancement has also been developed [10–13] in recent years. In [10] and [11], unreliable Mel-spectral components of noisy speech are imputed using a sparse linear combination of dictionary entries, and this sparse linear combination is computed based on reliable spectral components. The dictionary used in [10] is the discrete Haar transform basis; while in [11], the dictionary is made up of clean speech log-Mel-spectral training exemplars spanning several frames. The ASR performance using these missing data imputation techniques are found to be highly dependent on the accuracy of the binary mask that is used to decide the reliability of the spectral components. In [12], a dictionary containing clean speech and pure noise Mel-magnitude-spectral training exemplars are extracted from clean and noisy versions of the training data. Using all spectral components (both reliable and unreliable), the sparse linear combination of these Mel-spectral exemplars that best represents the test Mel-spectra, is derived by solving an L1-minimization problem. Subsequently, a soft-mask is estimated from exemplar-reconstructed clean and noisy spectra, which is used to denoise the test Mel-spectra prior to cepstral feature extraction. This feature enhancement technique outperforms the previous missing data imputation (with an estimated binary mask) scheme when evaluated on the Aurora-2 noisy digit speech recognition [14]. More recently in [13], another feature enhancement scheme is proposed using clean and noisy log-Mel-spectral exemplar-pairs that are extracted from clean and noisy utterance-pairs in the training data. The sparse linear combination of the noisy log-Mel dictionary exemplars that best approximates the log-Mel spectra of the noisy test speech is computed, and this sparse linear activation weighting vector is applied to the corresponding clean speech log-Mel dictionary exemplars to reconstruct the denoised log-Mel spectra. This technique reported a better ASR performance compared to the feature enhancement scheme in [12] on a small vocabulary, in-car noisy speech recognition task [15].

In this paper, feature enhancement for noisy speech recognition is performed using reference soft-mask (SM_{ref}) and estimated soft-mask (SM_{est}) exemplar-pairs. The SM_{ref} and SM_{est} exemplar-pairs are computed from clean and noisy utterance-pairs in the training data and stored in two separate dictionaries, D_r and D_e, respectively. The sparse linear combination of exemplars in D_e that best approximates the estimated soft-mask of the test utterance is computed by solving an L1-minimization problem. This sparse linear combination is then applied to the exemplars in D_r to generate an enhanced soft-mask, which is used to denoise the Mel-spectra before cepstral features are computed. The ASR performance of the proposed "Mask-2dict" technique on the Aurora-2 database is evaluated against existing sparse-exemplar-based feature enhancement techniques, which we abbreviate as "Mask-Mel-dict", "KL-Mask-Mel-dict" [12] and "LgMel-2dict" [13].

2. PROPOSED MASK-2DICT METHOD

2.1. Soft-mask exemplar dictionary generation

The joint SM_{ref} and SM_{est} exemplars dictionary generation scheme is shown in Fig. 1. The SM_{ref} of each time-frequency (T-F) unit is denoted by $M_r[k,t]$, where k and t are the Mel-frequency bin and time frame indices, respectively. In Eq (1), $M_r[k,t]$ is computed by taking the ratio of the clean Mel-frequency magnitude spectrum (Mel-spectrum), $Y_c[k,t]$, to the noisy Mel-spectrum, $Y_n[k,t]$, which is derived from the clean and noisy versions of the same training utterance, respectively. These Mel-spectra are computed by apply-

This work is supported in part by DARPA via SRI.

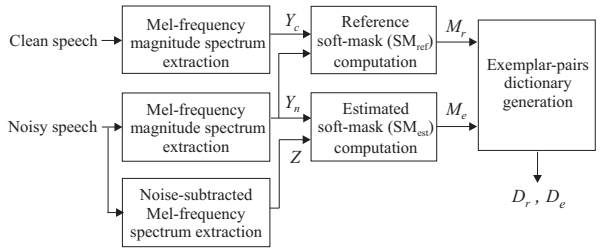

Fig. 1. Joint reference and estimated soft-mask exemplars dictionary generation scheme (training phase)

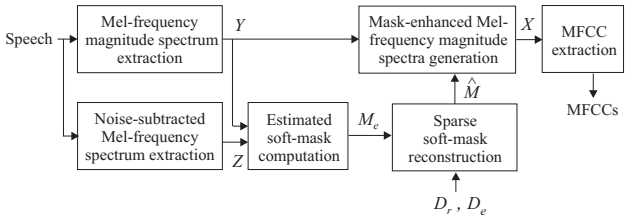

Fig. 2. The proposed feature enhancement algorithm using reference and estimated soft-mask exemplar-pair dictionaries (during MFCC extraction in both training and testing phases)

ing Mel-filter weighting on the respective pre-emphasized short-time FFT magnitude spectra. The SM_{est}, denoted by M_e, is computed as shown in Eq. (2), by taking the ratio of Z to Y_n, and flooring the resulting mask values to 0.05. Z is a denoised Mel-spectrum calculated by subtracting the estimated noise Mel-spectrum, $\hat{N}$ from Y_n, and negative values are set to 0 (Eq. (3)). $\hat{N}$ is obtained by applying Mel-filter weighting on the short-time FFT noise magnitude spectrum estimated from the pre-emphasized speech signal with the minimum statistics noise estimation technique [16] implemented in the "estnoisem" function of Voicebox [17]. Note that clean and noisy training utterance-pairs required for such exemplar-pairs extraction can be easily generated by introducing various noise or channel characteristics to clean training data.

$$M_r[k,t] = Y_c[k,t]/Y_n[k,t] \tag{1}$$

$$M_e[k,t] = \max(Z[k,t]/Y_n[k,t], 0.05) \tag{2}$$

$$Z[k,t] = \max(Y_n[k,t] - \hat{N}[k,t], 0) \tag{3}$$

The SM_{ref} and SM_{est} are extracted from all clean and noisy training utterance-pairs in the Aurora-2 database. To select a subset of the exemplars to build the dictionaries, we followed a similar random selection scheme described in [12]. We also tried the joint dictionary learning scheme in [13] with the referenced sparse coding toolbox [18] instead of the random selection scheme, but an insufficient memory issue arose during joint dictionary learning using an initial exemplar subset. Each exemplar-pair (m_r, m_e) is extracted from the same T-F region, and four exemplar-pairs are randomly selected from each clean and noisy training utterance-pair, where $m_* \in \mathbb{R}^{KT \times 1}$ is a column vector obtained by concatenating the columns in the respective mask T-F region $M_*[1{:}K, (\tilde{t}{+}1){:}(\tilde{t}{+}T)]$ that contains $K \times T$ elements, where K=23 is the total number of Mel-frequency bins, T=11 is the number of consecutive frames in the Mel-spectrum (which approximately spans a phoneme interval), and $\tilde{t} \in [0, \text{total frames - } T]$ is a randomly generated frame index. The notation $a{:}b$ represents the range of integers $\{a, a{+}1, \ldots, b\}$. T=11 is also used in [13], and good ASR performance is achieved by the feature enhancement scheme in [12] using T=10. From this initial subset, 4000 exemplar-pairs ($\tilde{m}_r^i$,$\tilde{m}_e^i$), i = 1, 2, ... , 4000, are randomly selected to form the dictionary. We selected 4000 pairs because 8000 dictionary exemplars (4000 from clean speech, 4000 from pure noise) are found to be sufficient in [12] for the Aurora-2 noisy digit ASR task. The L2 norm, L_e^i, of each $\tilde{m}_e^i$ exemplar is calculated, which is then used to normalize both $\tilde{m}_e^i$ and $\tilde{m}_r^i$ exemplars to form the respective SM_{est} and SM_{ref} dictionaries, D_e and D_r, as shown in Eqs. (4) and (5).

$$D_e = [d_e^1, d_e^2, ..., d_e^{4000}], \text{ where } d_e^i = \tilde{m}_e^i / L_e^i \tag{4}$$

$$D_r = [d_r^1, d_r^2, ..., d_r^{4000}], \text{ where } d_r^i = \tilde{m}_r^i / L_e^i \tag{5}$$

2.2. Feature enhancement

Fig. 2 summarizes the proposed feature enhancement scheme that uses the joint SM_{ref} and SM_{est} dictionaries. The SM_{est} of the input speech utterance, M_e, is computed in the same way as described in Section 2.1. M_e is divided into overlapping ($K \times T$) T-F regions, at 1 frame shift apart. The soft-mask vector starting from frame j is denoted by m_e^j, which is obtained by concatenating columns in $M_e[1{:}K, j{:}(j{+}T{-}1)]$. The sparse linear combination of D_e exemplars that best reconstructs each m_e^j is computed by solving the L1-minimization problem in Eq. (6), known as Lasso [19]. The sparse vector x^j is calculated via the *SolveLasso* function in the SparseLab toolbox [20]. The value of λ in Eq. (6) is iteratively updated in SparseLab's Lasso implementation. The "nnlasso" (non-negative Lasso) algorithm is used, and maximum number of iterations is set to 50. Default values are used for other input parameters to the *SolveLasso* function. The enhanced soft-mask vector, $\hat{m}^j$ is obtained by multiplying the same sparse vector, x^j to D_r (Eq. (7)).

$$\min_{x^j} \lambda ||x^j||_1 + 0.5||m_e^j - D_e x^j||_2^2 \quad \text{s.t. } x^j \geq 0 \tag{6}$$

$$\hat{m}^j = D_r\, x^j \tag{7}$$

After $\hat{m}^j$ are obtained for all frames, they are shaped back into rectangular T-F regions at their original frame positions, i.e. $\hat{M}[1{:}K, j{:}(j{+}T{-}1)]$. Overlapping T-F units are averaged [12] to obtain the enhanced soft-mask of the entire test utterance $\hat{M}[k,t]$. The enhanced Mel-spectrum of the test utterance, X, is then obtained by multiplying the derived enhanced soft-mask $\hat{M}$ with the original Mel-spectrum, Y, as shown in Eq. (8). Logarithmic magnitude compression is applied on X, followed by discrete Cosine transform (DCT) and liftering to obtain the first 13 cepstral coefficients (C0–C12). These are concatenated with their deltas and double-deltas to form the 39-dimension MFCC feature vector for ASR.

$$X[k,t] = \hat{M}[k,t]\, Y[k,t] \tag{8}$$

3. ALGORITHMS FOR COMPARISON

3.1. LgMel-2dict

This is based on [13], in which clean and noisy speech log Mel-spectrum dictionary exemplar-pairs are used, instead of the soft-mask exemplars in the proposed algorithm. Dictionary construction follows the procedure described in Section 2.1, even exemplars from the same 4000 ($K \times T$) T-F patches are selected to form the clean speech dictionary (D_c), and noisy speech dictionary (D_n). The sparse linear combination of D_n exemplars that best represents the input speech's log Mel-spectrum is found by solving the same L1-minimization problem with the same *SolveLasso* configuration. Similarly, this sparse solution is multiplied to D_c to reconstruct a denoised log Mel-spectrum, from which MFCCs are extracted.

3.2. Mask-Mel-dict

This is adapted from [12], in which Mel-spectrum exemplars of clean speech and pure noise are extracted. Noise signals are obtained from the noisy speech utterances by subtracting the clean speech signal (read from the clean version of the same training utterance)

from the noisy speech signal. The same 4000 ($K \times T$) T-F patches are selected to form two dictionaries, D_s and D_n, which contains 4000 Mel-spectrum exemplars from clean speech and pure noise, respectively. Each exemplar is normalized to have an L2 norm of 1. These two dictionaries are combined to form $D = [D_s, D_n]$, and the sparse linear combination of exemplars in D that best represents the input speech's Mel-spectrum (Y) is found by solving the same L1-minimization problem. After which, the sparse weights corresponding to the clean speech and noise exemplars are used to construct the clean speech Mel-spectrum (S) and noise Mel-spectrum (N), respectively. A soft-mask is then computed by taking the ratio $S/(S+N)$, which is multiplied to Y to obtain the denoised Mel-spectrum before MFCC extraction.

3.3. KL-Mask-Mel-dict

This closely follows the feature enhancement algorithm in [12] in the computation of the soft-mask for Mel-spectrum enhancement before MFCC extraction. The three main differences between this implementation and Mask-Mel-dict are (1) both the rows and columns of the dictionary are normalized, (2) a higher sparsity parameter is used for speech exemplars than noise exemplars to make the activation weights of speech exemplars more sparse, and (3) the generalized Kullback-Leibler (KL) divergence replaces the Euclidean distance in Eq. (6)'s L1-minimization expression (refer to Eq. (8) in [12]), which is run over 200 iterations. To reduce the run-time, we stop the iterative update when $||x_k - x_{k-1}||_2/||x_k||_2 < 0.01$, where x_k and x_{k-1} are the sparse solutions obtained in the current and previous iterations, respectively.

4. DATABASE AND EXPERIMENTAL SETUP

The Aurora-2 noisy digit speech recognition task is used to evaluate the performance of the feature enhancement algorithms. Feature enhancement is applied to MFCC extraction during both training and testing. The standard HMM architecture [14] is built using the HTK software package [21]. The database contains two training sets – (1) clean, (2) multi-conditional. The multi-conditional training set contains ITU G.712-filtered [22] clean speech and noisy speech (SNRs between 20 and 5 dB). Suburban train, babble, car, or exhibition hall noise is artificially added to clean speech to generated each noisy utterances in the multi-conditional training set. Since the exemplar dictionaries are already computed with knowledge of the noisy training utterances, the HMMs are trained using the multi-conditional set. There are three testing sets – (1) Test A, (2) Test B, and (3) Test C. Test A and Test B contains G.712-filtered speech (SNRs between 20 to -5 dB). The noise-types in Test A are the same as those found in the multi-conditional training set, while a different set of noise-types (restaurant, street, airport, train-station) is found in Test B. Test C contains ITU MIRS-filtered [22] utterances, and the noise-types involved are suburban-train and street noises. The major difference between the frequency characteristics of G.712 and MIRS filters is that the former has a flat response in the range between 300 and 3400 Hz, while the latter has a rising response with a greater attenuation at lower frequencies (illustrated in Fig. 1 of [14]). MIRS simulates the telecommunication terminal input frequency response in the technical specification GSM 03.50 [23].

5. RESULTS

Tables 1 and 2 contain the word accuracy (Acc) results obtained using models trained with the multi-condition training set, where the latter results are obtained with an additional feature mean and variance normalization (MVN) performed on a per utterance basis. We also include the results obtained using MFCCs for comparison.

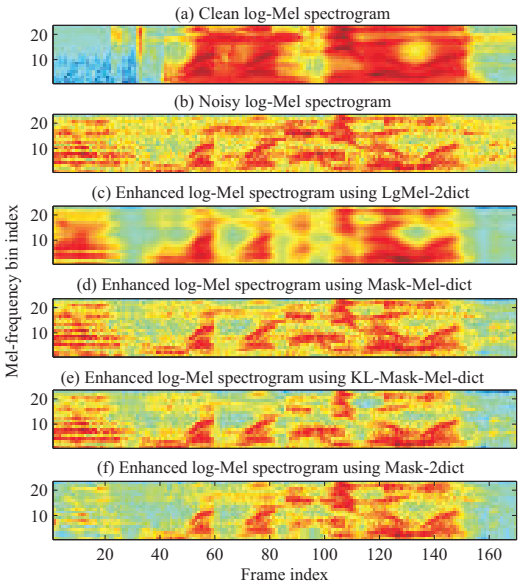

Fig. 3. Log-Mel spectrograms of a test utterance corrupted by airport noise at 0 dB SNR. (a) Clean log-Mel spectrogram, (b) Noisy log-Mel spectrogram, (c)–(f) Enhanced log-Mel spectrograms obtained using the lgMel-2dict, Mask-Mel-dict, KL-Mask-Mel-dict, and Mask-2dict, respectively.

From Tables 1 and 2, it can be observed that the Mask-Mel-dict method has the best performance, on average, for Test A. However, for Tests B and C, the performance of sparsity-based methods (LgMel-2dict, Mask-Mel-dict, KL-Mask-Mel-dict) that utilize spectral-based dictionaries decreases sharply, such that their Accs are lower than those obtained by MFCCs in some cases -- – LgMel-2dict performance is poorer than MFCCs with/without MVN, while Mask-Mel-dict and KL-Mask-Mel-dict achieve lower Accs than MFCC when MVN is applied. In contrast, the proposed Mask-2dict technique that uses joint SM_{ref} and SM_{est} exemplars has the best performance among the other sparsity-based methods for Tests B and C, with significant gains in Acc over these methods at low SNRs.

Fig. 3 plots the denoised log-Mel spectrograms obtained by the sparse exemplar-based techniques evaluated in this study for a test utterance corrupted by airport noise at 0 dB SNR (this noise-type is not present in the multi-conditional training set). In this example, the proposed technique does a better job in noise suppression at the beginning of the utterance.

6. DISCUSSION

Comparative methods using Mel-spectral exemplars, whether with dual (or joint) dictionaries (in the case of LgMel-2dict), or with a combined dictionary (in the case of Mask-Mel-dict, KL-Mask-Mel-dict), perform well when the test noise and channel frequency characteristics match those present in the dictionary, as observed for Test A. However, the performance of these methods for Tests B and C suffers a large degradation when spectral shape mismatches are present. On the other hand, the proposed technique using the dual soft-mask exemplar dictionaries, is less sensitive to such spectral shape mismatches, with less performance degradation observed across the three test sets. One possible reason is that the estimated

Table 1. Acc (%) obtained on Aurora-2 using the multi-condition training set. The "Avg." value is computed by averaging the Acc over SNRs between 20 and 0 dB. The highest Acc among all algorithms are in bold.

Algorithm	Test A				Test B				Test C			
	20 dB	10 dB	0 dB	Avg.	20 dB	10 dB	0 dB	Avg.	20 dB	10 dB	0 dB	Avg.
MFCC	97.73	94.50	55.71	86.27	97.12	92.89	60.18	86.08	96.82	92.95	49.69	83.73
LgMel-2dict	97.68	94.50	60.95	87.55	96.87	91.54	53.45	83.58	96.03	86.95	32.38	75.29
Mask-Mel-dict	98.22	**95.88**	**79.22**	**92.46**	98.15	93.36	61.13	86.63	**97.96**	93.04	55.60	85.02
KL-Mask-Mel-dict	**98.33**	95.84	77.26	92.08	**98.33**	92.93	60.27	86.23	97.66	92.19	53.71	84.12
Mask-2dict	98.30	95.78	71.52	90.82	98.27	**94.52**	**67.71**	**89.01**	97.95	**94.69**	**64.18**	**88.21**

Table 2. Acc (%) obtained on Aurora-2 using the multi-condition training set, with mean and variance normalization (MVN) applied.

Algorithm	Test A				Test B				Test C			
	20 dB	10 dB	0 dB	Avg.	20 dB	10 dB	0 dB	Avg.	20 dB	10 dB	0 dB	Avg.
MFCC	98.53	95.99	72.73	91.03	**98.57**	96.21	73.24	91.29	98.48	95.20	72.50	90.59
LgMel-2dict	97.83	95.72	74.09	91.06	97.80	95.03	66.21	88.71	97.32	91.69	56.99	84.61
Mask-Mel-dict	98.64	**96.75**	**80.73**	**93.31**	98.55	95.88	69.29	90.07	**98.51**	94.70	64.50	88.31
KL-Mask-Mel-dict	98.53	96.38	80.19	93.03	98.50	95.65	69.72	90.05	98.33	94.90	63.60	88.21
Mask-2dict	**98.70**	96.57	75.25	91.95	98.56	**96.54**	**74.34**	**91.74**	98.28	**95.84**	**75.37**	**91.50**

Table 3. Avg. Acc (%) obtained on 8-kHz Aurora-4 test data using multi-noise training and MVN

Algorithm	Seen noise-types		Unseen noise-types	
	Babble	Street	F16	Pink
Mask-Mel-dict	**77.66**	**76.48**	70.05	59.89
Mask-2dict	76.65	75.81	**75.55**	**67.98**

mask is computed using a noise estimation algorithm that does not make any assumption regarding the noise spectral shape present in the utterance. We also ran the ASR experiment using enhanced MFCCs obtained by directly applying the estimated soft-mask, M_e, to the noisy Mel-spectrum. The average Acc obtained with MVN on test sets A, B, and C are 88.66%, 89.65% and 86.74%, respectively, which are 3–5% worse than those achieved with the proposed algorithm. This shows that the sparse mask reconstruction step is essential in enhancing ASR performance, and it can potentially improve with more accurate noise estimation. We are aware that supplementing the dictionary with artificial noise exemplars or noise exemplars extracted from the initial frames of test utterance [24], and channel compensation [25] can improve the performance of Mask-Mel-dict and KL-Mask-Mel-dict on mismatched noise and channel conditions. However, in this study, we are evaluating the algorithms' performance on exemplars extracted solely from the training data, without performing explicit channel compensation.

To assess if similar performance trends hold in a larger vocabulary setup, a preliminary experiment was run on the Aurora-4 (A4) database [26], with dictionary exemplars derived from A4 clean and noisy training sets. Besides evaluating on two A4 test sets (each containing one of the six noise-types in the multi-noise training set), we added two unseen noise-types from Noisex-92 [27] (not in the multi-noise training set) to the clean A4 test set to simulate Test B scenario. ASR is performed on 8-kHz files recorded with the Sennheiser microphone, using word-internal triphone models, multi-noise training (same microphone) and MVN. Similar Acc trends are observed in Table 3 – Mask-Mel-dict Acc's are higher than Mask-2dict's for seen noise-types, and vice versa for unseen noise-types.

The advantage of using a soft-mask as part of feature enhancement can also be observed by comparing the performance of LgMel-2dict with the other two spectral-based dictionary methods (Mask-Mel-dict and KL-Mask-Mel-dict). Generating denoised spectra by applying a soft-mask on the original noisy spectra tends to be more error-forgiving compared to reconstructing it from clean spectral exemplars as done in the LgMel-2dict method. We also observe a decrease of 2–3 % in absolute Avg. Acc for all test sets when the denoised Mel-spectra (S) is reconstructed directly from the sparse linear combination of clean exemplars (results of this variant implementation are not shown in the Tables), instead of reconstructing it indirectly via the soft-mask in the Mask-Mel-dict method.

We implemented all the sparse-exemplar-based techniques using MATLAB. LgMel-2dict takes the shortest time for feature extraction at ≈4.3×real-time (RT) on a Intel Xeon 2.6 GHz processor (parallel-core computing is not utilized), followed by the proposed Mask-2dict at ≈4.7×RT due to the additional noise estimation step. Mask-Mel-dict takes ≈11×RT due to double the number of dictionary exemplars used to compute the sparse activation vector, and KL-Mask-Mel-dict takes ≈28×RT due to a large number of iterations (200) used to perform the KL divergence L1-minimization.

We intend to explore other dictionary training packages for constructing the sparse representation dictionaries. This could help in extracting more representative exemplars over random selection for large vocabulary ASR, and also potentially improve the performance of joint dictionary schemes through joint dictionary training.

7. CONCLUSION

A novel feature enhancement scheme using sparse reference soft-mask (SM_{ref}) and estimated soft-mask (SM_{est}) exemplar-pairs is proposed. SM_{ref} is the ratio of clean Mel-spectrum to the noisy Mel-spectrum computed from a clean and noisy utterance-pair in the training data, while SM_{est} is the ratio of a denoised Mel-spectrum to the original noisy Mel-spectrum. The denoised Mel-spectrum is obtained by subtracting the noise spectrum (estimated using the minimum statistics noise estimation algorithm) from the original noisy spectrum. The sparse linear combination of SM_{est} dictionary exemplars that best approximates the SM_{est} of the test speech utterance is found by solving an L1-minimization problem. An enhanced soft-mask is generated by applying the same sparse linear combination to the SM_{ref} dictionary exemplars. This soft-mask is used to enhance the Mel-spectra before MFCCs are extracted. Compared to other existing sparse-exemplar-based feature enhancement techniques that utilizes Mel-spectral-based dictionary exemplars, the proposed scheme achieves higher word accuracies in the presence of spectral shape mismatch between dictionary exemplars and the test set, when evaluated on the Aurora-2 ASR task with multi-conditional training. Similar trends are observed in a preliminary experiment on Aurora-4.

8. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1979, pp. 208–211.

[2] J. van Hout and A. Alwan, "A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition," in *IEEE ICASSP*, 2012, pp. 4105–4108.

[3] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE ICASSP*, 2013, pp. 7092–7096.

[4] S. Y. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition.," in *INTERSPEECH*, 2008, pp. 898–901.

[5] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *IEEE ICASSP*, 2012, pp. 4117–4120.

[6] M. J. F Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, pp. 75–98, 1998.

[7] A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang, "HMM adaptation using vector taylor series for noisy speech recognition," in *INTERSPEECH*, 2000, pp. 869–872.

[8] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 210–227, 2009.

[9] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[10] B. J. Borgstrom and A. Alwan, "Utilizing compressibility in reconstructing spectrographic data, with applications to noise robust ASR," *IEEE Signal Processing Letters*, vol. 16, pp. 398–401, 2009.

[11] J. F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 272–287, 2010.

[12] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2067–2080, 2011.

[13] W. Li, Y. Zhou, N. Poh, F. Zhou, and Q. Liao, "Feature denoising using joint sparse representation for in-car speech recognition," *IEEE Signal Processing Letters*, vol. 20, pp. 681–684, 2013.

[14] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA Tutorial and Research Workshop (ITRW) for Automatic Speech Recognition: Challenges for the new Millenium*, 2000, pp. 181–188.

[15] M. Fujimoto, K. Takeda, and S. Nakamura, "CENSREC-3: An evaluation framework for Japanese speech recognition in real car-driving environments," *IEICE Transactions on Information and Systems*, vol. 89, pp. 2783–2793, 2006.

[16] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, pp. 1215–1229, 2006.

[17] M. Brookes, "Voicebox: Speech Processing Toolbox for MATLAB," 2011, Software available, URL: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[18] H. Lee, A. Battle, R. Raina, and A. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems*, 2006, pp. 801–808, Software available, URL: http://ai.stanford.edu/~hllee/softwares/nips06-sparsecoding.htm.

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

[20] D. L. Donoho, V. C. Stodden, and Y. Tsaig, "About SparseLab (version 2.1)," 2007, Software available, URL: http://sparselab.stanford.edu/.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, XA Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, 2006.

[22] ITU Recommendation, "G.712: Transmission performance characteristics of pulse code modulation channels," 1996.

[23] ETSI-GSM technical specification, "European digital cellular telecommunication system (Phase 1); Transmission planning aspects for the speech service in GSM PLMN system," GSM 03.50, version 3.4.0, 1994.

[24] J. F. Gemmeke and H. van Hamme, "Advances in noise robust digit recognition using hybrid exemplar-based techniques," in *INTERSPEECH*, 2012, pp. 2134–2137.

[25] J. F. Gemmeke, T. Virtanen, and K. Demuynck, "Exemplar-based joint channel and noise compensation," in *IEEE ICASSP*, 2013, pp. 868–872.

[26] N. Parihar, J. Picone, D. Pearce, and H. G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," in *Eurospeech*, 2004, pp. 553–556.

[27] A. Varga and H. 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 Communication*, vol. 12, pp. 247–251, 1993.