SEMI-SUPERVISED NOISE DICTIONARY ADAPTATION FOR EXEMPLAR-BASED NOISE ROBUST SPEECH RECOGNITION

Yi Luan† *Daisuke Saito*† *Yosuke Kashiwagi*† *Nobuaki Minematsu*† *Keikichi Hirose*†

† The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo 113-8656 Japan

ABSTRACT

The exemplar-based approaches, which model signals as a sparse linear combination of exemplars of signals, are proved to have state-of-the-art performance in noise robust ASR, especially on low SNRs. However, since both the speech exemplars and noise exemplars are built from training data and are fixed throughout the process of enhancing speech features, the conventional approach is especially weak for unknown types of noise. Therefore, in this paper, we propose a semi-supervised approach which automatically adapt noise exemplars to the target noise, while keeping the speech exemplars fixed. Continuous digits recognition experiments show that this approach is much more robust for unknown noise. The recognition errors are reduced by 36.2%.

Index Terms— robust speech recognition, non-negative matrix factorization, semi-supervised, exemplar-based, noise reduction

1. INTRODUCTION

The Automatic Speech Recognition (ASR) technology now has been proved to have desirable performance in an environment that is exactly the same as that observed in training the recognition model. The HMM is the most popular and successful stochastic approach to speech recognition in general use, due to the existence of elegant and efficient algorithms for both training and recognition. However, in the actual world, it is very hard to predict what kind of acoustic condition the speakers speak in. Therefore, background noise, channel distortion and speaker variations sometimes show great mismatch between the training and testing conditions, which often causes a dramatic degradation in performance of the ASR systems.

There have been numerous approaches that aim at reducing this mismatch. The approaches can be split into three categories. Firstly, inherently robust parametrization of speech may be used, such as Histogram Equalization and Linear Discriminant Analysis [1]. Alternatively, the clean speech may be estimated from its corrupted version, then the clean acoustic models may be used as they are. The famous approaches are Spectral Subtraction [2], Stereo-based Piecewise Linear Compensation for Environments [3] and Vector Taylor Series approaches [4]. Finally, the models can be made to fit the testing condition. For example, Maximum Likelihood Linear Regression (MLLR) uses a set of linear transforms to adapt the model parameters to a new acoustic condition.

Recently a class of methods that has gained recent prominence is based on compositional models: noisy speech spectra are represented as a linear, typically sparse, combination of basis atoms describing the individual speech and noise sources. The collection of atoms here is called a dictionary, which includes both clean speech exemplars and noise exemplars [5]. The sparse representation is obtained by finding the sparsest possible linear combination that describes the observed signal well, using techniques best known as non-negative matrix factorisation (NMF) [6]. Reconstruction of an observed spectrum as weighted sum of parts of the dictionary can be used to separate the spectrum into clean and noisy spectra and the clean spectrum will be used for recognition.

However, in previous studies[5], since the noise exemplar in the dictionary is fixed, for unexpected types of noise, the performance of separating into speech and noise would be greatly decreased. If acoustic mismatch exists in the noise type between the exemplars and testing data, the speech parts of testing data will not be reconstructed well. In order to solve the problem, we propose a semi-supervised method that could adapt the noise exemplars together with the activation matrix while remaining the speech exemplars supervised. This semi-supervised exemplar adaptive method is more flexible because it can absorb the various type of noise. AURORA-2 is utilized for testing the method. Our semi-supervised approach is proved to show much better performance than conventional approach for test set B in AURORA-2, which has an open noise condition. The average recognition rate for supervised NMF is 78.51, while our semi-supervised approach gains much higher recognition rate which is 84.22 (36% relative increase).

The remainder of this paper is organized as follows. In Section 2, we give a brief review of previous studies of supervised exemplar approach. In Section 3, the proposed noise exemplar update method is introduced and explained. The results of experiments are described in Section 4. Finally Chapter 5 draws overall conclusions and describes possible future

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

work.

2. SUPERVISED EXEMPLAR-BASED APPROACH

NMF is an useful decomposition method for multivariate data. Each testing data vector can be approximated by a linear combination of the basis, weighted by the activation. In speech processing, the noisy speech can be represented in the form of NMF as is shown in equation (1).

The magnitude spectrogram describing a whole speech segment is a $B \times T_s$ dimensional matrix. B is the number of mel-frequency bands of the speech, T_s is the number of samples in that utterance. In order to decompose the utterance magnitude spectrogram of length T_s, a sliding window approach is adopted as in [5]. An utterance is divided into a number of overlapping, fixed-length windows, with window length of T_{win}. The columns of in each window are stacked into a single vector of length $W = B \times T_{win}$. Through applying sliding window, the original magnitude spectrogram can be transformed to $\boldsymbol{y}$, which is a $W \times N$ matrix, where N is the total number of sliding windows in the utterance.

$$\boldsymbol{y} \approx \boldsymbol{s} + \boldsymbol{n} \tag{1}$$

$$\approx \sum_{j=1}^{J} \boldsymbol{a}_j^s \boldsymbol{x}_j^s + \sum_{k=1}^{K} \boldsymbol{a}_k^n \boldsymbol{x}_k^n \tag{2}$$

$$= \begin{bmatrix} \boldsymbol{A}^s & \boldsymbol{A}^n \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^s \\ \boldsymbol{x}^n \end{bmatrix} s.t.\ \boldsymbol{x}^s, \boldsymbol{x}^n \geq 0 \tag{3}$$

$$= \boldsymbol{A}\boldsymbol{x} \tag{4}$$

$\boldsymbol{s}$ and $\boldsymbol{n}$ are the clean and noise parts of the utterance spectrogram respectively. $\boldsymbol{A}^s$ and $\boldsymbol{A}^n$ are called the clean dictionary and noise dictionary which are the matrices containing clean speech exemplars $\boldsymbol{a}_j^s$ and noise exemplars $\boldsymbol{a}_k^n$, with the dimension of $W \times J$ and $W \times K$, respectively. The whole exemplar dictionary has the number of $L = J + K$ exemplars. Each exemplar is one slide window extracted randomly from the training data and is stacked to a W length vector. $\boldsymbol{x}^s$ and $\boldsymbol{x}^n$ are the activation vectors of the clean speech and noise exemplars, with dimension of $J \times N$ and $K \times N$. The requirement of $\boldsymbol{x}$ is non-negative and sparse.

To obtain sparse representation vector $\boldsymbol{x}$, we use the following cost function to minimize.

$$d(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}) + ||\boldsymbol{\lambda}. * \boldsymbol{x}||_p \tag{5}$$

The first term d measures the distance between the noisy observation and its approximation based on NMF. The second term enforces sparsity by penalizing the non-zero entries of $\boldsymbol{x}$ weighted by $\boldsymbol{\lambda}$. Kullback-Leibler(KL) divergence is used for d here, as (6). y_e and $\hat{y}_e$ are the elements of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_e$.

$$d(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{e=1}^{E} y_e \log(\frac{y_e}{\hat{y}_e}) - y_e + \hat{y}_e \tag{6}$$

Enforcing the sparseness of speech exemplars is much more important than noise exemplars, since an observed speech segment should be ideally represented only by adding up a part of atoms of the speech dictionary. We do not make such assumptions about noise and thus do not enforce the sparseness of noise exemplars.

The cost function (5) is minimized by firstly initializing the vector $\boldsymbol{x}$ to unity, and then iteratively applying the update rule:

$$x \leftarrow x. * (\boldsymbol{A}^T(\boldsymbol{y}./(\boldsymbol{A}\boldsymbol{x}))./(\boldsymbol{A}^T\boldsymbol{1} + \boldsymbol{\lambda}) \tag{7}$$

$.*$ and $./$ denote element-wise multiplication and division, respectively. The vector $\boldsymbol{1}$ is an all-one vector of length W. The derivation of (7) is given in [5].

3. SEMI-SUPERVISED NMF APPROACH

The supervised approach has drawbacks when there is a mismatch between the noise exemplars and noise input in test data. The number of noise exemplars can be numerous in order to include all types of noise, which may result in huge computational cost. Therefore, we propose a semi-supervised approach to solve this problem.

Instead of using fixed noise exemplars, we propose to update noise exemplars together with updating the activation matrix. The motivation of doing this modification is that we would like the noise exemplars to be more flexible, so that it can handle various and different kinds of noise automatically. However, if we use cost function like (5), the effect of updated noise exemplars would become so strong that it would even influence the $\boldsymbol{x}_s$ inadequately. Therefore, we modify the cost function as follows,

$$d(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}) + ||\boldsymbol{\lambda}. * \boldsymbol{x}||_p + \boldsymbol{\eta}. * d(\boldsymbol{N}, \hat{\boldsymbol{A}}_n). \tag{8}$$

The third term of cost function in (8) represents the KL divergence from the updated noise exemplars $\hat{\boldsymbol{A}}_n$ to the original noise exemplars $\boldsymbol{N}$. The purpose of adding the third term is to suppress the effect of updating noise exemplars. The update rule for (8) is as follows,

$$x \leftarrow x. * (\boldsymbol{A}^T(\boldsymbol{y}./(\boldsymbol{A}\boldsymbol{x}))./(\boldsymbol{A}^T\boldsymbol{1} + \boldsymbol{\lambda}), \tag{9}$$

$$\boldsymbol{A}_n \leftarrow (\boldsymbol{A}_n.*(\boldsymbol{y}./(\boldsymbol{A}\boldsymbol{x})*\boldsymbol{x}_n^T) + \boldsymbol{\eta}.*\boldsymbol{N})/(\boldsymbol{1}*\boldsymbol{x}_n^T + \boldsymbol{\eta}). \tag{10}$$

The derivation of (9) and (10) is given in appendix.

4. EXPERIMENTS

We use AURORA-2 to test the performance of semi-supervised approach. Experimental condition is shown in Table 1.

Acoustic feature vectors used in the exemplar-based framework consisted of Mel-frequency magnitude spectra: 23 frequency bands plus log-energy, $B = 23 + 1$ in total. The exemplar size for the speech and noise dictionaries is $T_{win} = 20$ frames. The speech exemplars are randomly

Table 1. Experimental Condition

GMM mixture	256
Exemplar numbers for supervised approach	4000 for speech, 4000 for noise
Exemplar numbers for semi-supervised approach	4000 for speech, 2 for noise
Feature for decoding	MFCC_E_D_A
training data	clean condition in AURORA-2
test data	Test set A& Test set B in AURORA-2

selected from training data set. The speech exemplars keep fixed both in supervised and semi-supervised experiments. In the experiment for supervised approach, the noise exemplars are randomly selected from the noise data set in training data set (from SNR5 to SNR20). In the experiment for semi-supervised approach, to prevent over-fitting problems, only two exemplars are used. The noise exemplars $\hat{\boldsymbol{A}}_n$ are initialized by $\mathbf{1}$, and are updated together with the activation matrix. This initialization is proved to have better performance than randomly select 2 noise exemplars in training data set, since it gives equal initialization for each mel-frequency band. The penalizing weight $\boldsymbol{\lambda}$ for supervised approach is 0.65 for clean exemplars and 0 for noise exemplars. The penalizing weight $\boldsymbol{\lambda}$ for semi-supervised approach is 0.65 for clean exemplars and 0.5 for noise exemplars, while $\boldsymbol{\eta} = 1$. $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ are chosen by using 2-fold cross validation.

Test sets A and B in AURORA-2 are used for testing. Test set A has the close noise condition which has the same type of noise as the training data, while Test set B has open noise condition which has completely different noise type as in the training data. The clean condition training data set is used for both training acoustic model for recognition and for building the speech dictionary. In order to reduce mismatch between the acoustic model and the testing data, we enhance both the training data and the testing data using the same NMF enhancement approach.

The results are shown in Table 2 and Table 3. For Test set A there is no significant improvement when semi-supervised approach is applied. For Test set B, since the type of noise is different from training data, the mismatch between the noise exemplars and the test data causes great performance degradation of supervised-approach. For semi-supervised approach, since no empirical value of noise is included in the exemplar dictionary, the recognition rate got great improvement with 36% relative increase than the supervised approach. In order to discover the effect of semi-supervised approach on each specific type of noise, the different types of noise and recognition rate in Test set A and B are listed in detail as Table 4.

From Table 4, when human noise is included in the test data, semi-supervised approach would find it difficult to separate the speech and noise components in the noisy data. Since no prior knowledge of noise is included in the exemplar dictionary, NMF would classify all the speech elements together in a situation when human noise are included. In Test set A, babble noise are pure human noise, which results in worst recognition rate. In Test set B, human noise exists in both restaurant and airport condition, which leads to lower recognition rate of these two sets. The situation is also true to supervised approach due to the close characteristic between babble noise and clean speech, which means exemplar approaches are not good at reducing human noise.

5. SUMMARY AND CONCLUSION

Semi-supervised approach is proved to be much more robust than supervised approach for unknown noise conditions. The noise exemplars in semi-supervised approach is very flexible to absorb noise with variant characteristics, in regardless of the prior knowledge of noise in training data. The semi-supervised approach gains 36% relative increase than supervised approach in Test set B, which is an open noise condition data set in AURORA-2 database. Meanwhile, we observe that

Table 2. Recognition rate of Test A

SNR	Supervised	Semi-supervised
clean	98.96	99.08
SNR20	96.00	97.99
SNR15	93.65	95.99
SNR10	89.35	90.37
SNR5	81.47	80.36
SNR0	65.65	62.66
SNR-5	42.07	41.07
Average	85.24	85.48

Table 3. Recognition rate of Test B

SNR	Supervised	Semi-supervised
clean	98.96	99.08
SNR20	97.39	98.05
SNR15	94.18	95.44
SNR10	84.98	89.32
SNR5	68.32	77.51
SNR0	47.62	60.78
SNR-5	28.00	40.92
Average	**78.51**	**84.22**

Table 4. Average recognition rate of Test set A and Test B specified on types of noise

Test set A	Subway	Babble	Car	Exibition
Average	88.97	77.97	89.00	85.97
Test set B	Restaurant	Street	Airport	Train Station
Average	80.44	86.27	83.42	86.76

the recognition rate decreases when human noise is included in the test data for both supervised and semi-supervised approach, due to the similarity between the human noise and the speech. In our future work, we would try to solve this problem using exemplar approach by combining some other feature enhancement approaches.

6. APPENDIX: DERIVATION OF THE UPDATE RULE

The update rules (9) and (10) are derived from the auxiliary function of the first term of (8) by the similar manner to the ordinary NMF [6]. The upper bound of $d(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x})$ is derived as follows:

$$
\begin{aligned}
& d(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}) \\
&= \sum_{w,t} \Big(y_{w,t} \log \frac{y_{w,t}}{\sum_{l=1}^{L} a_{w,l} x_{l,t}} - y_{w,t} + \sum_{l=1}^{L} a_{w,l} x_{l,t} \Big) \\
&= \sum_{w,t} (y_{w,t} \log y_{w,t} \underline{- y_{w,t} \log \sum_{l=1}^{L} a_{w,l} x_{l,t}} - y_{w,t} + \sum_{l=1}^{L} a_{w,l} x_{l,t}) \\
&\leq \sum_{w,t} (y_{w,t} \log y_{w,t} \underline{- y_{w,t} \sum_{l=1}^{L} \gamma_l \log \frac{a_{w,l} x_{l,t}}{\gamma_l}} - y_{w,t} + \sum_{l=1}^{L} a_{w,l} x_{l,t}) \\
&\doteq Q(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}, \gamma) \ \text{ s.t.} \sum_{l} \gamma_l = 1. \qquad (11)
\end{aligned}
$$

In the above derivation, Jensen's inequality is used for the underlined term. The condition of equality is satisfied when

$$
\hat{\gamma}_l = \frac{a_{w,l} x_{l,t}}{\sum_{l}^{L} a_{w,l} x_{l,t}}. \qquad (12)
$$

Therefore, the upper bound function of (8) that should be minimized is

$$
Q(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}, \hat{\gamma}) + ||\boldsymbol{\lambda}. * \boldsymbol{x}||_p + \boldsymbol{\eta}. * d(N, \hat{\boldsymbol{A}}_n). \qquad (13)
$$

Setting the derivative of (**??**) with respect to $\boldsymbol{A}$ and $\boldsymbol{x}$, we finally derive the following update rules for each element:

$$
a_{w,j} = \frac{\sum_t y_{w,t} \frac{a_{w,j} x_{j,t}}{\sum_l a_{w,l} x_{l,t}}}{\sum_t x_{l,t}} \qquad (14)
$$

$$
a_{w,n} = \frac{\sum_t y_{w,t} \frac{a_{w,n} x_{n,t}}{\sum_l a_{w,l} x_{l,t}} + \eta N_{w,n}}{\sum_t x_{n,t} + \eta} \qquad (15)
$$

$$
x_{l,t} = \frac{\sum_w y_{w,t} \frac{a_{w,l} x_{l,t}}{\sum_l a_{w,l} x_{l,t}}}{\sum_w a_{w,l} + \lambda} \qquad (16)
$$

where $a_{w,j}$ is the element in speech exemplars, while $a_{w,n}$ is the element in noise exemplars. $x_{l,t}$ refers to activation matrix. Therefore, we can get the updating rules (9) and (10).

7. REFERENCES

[1] M. J. Hunt and C. Lefebre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. ICASSP*, 1989.

[2] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. of IEEE*, 1995.

[3] J. Droppo and L. Deng, "Evaluation of SPLICE on the aurora 2 and 3 tasks," in *Proc. ICSLP*, 2002, pp. 29–32.

[4] M. J. F. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," in *Proc. Computer Speech & Language*, 2010.

[5] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing , IEEE Trans.*, 2011.

[6] D. Lee and H. S. Seung, "Algorithms for non-negative matrix fac- torization," in *Proc. NIPS*, 2000.