NOISY SPEECH RECOGNITION USING TEMPORAL AM-FM COMBINATION

Yotaro KUBO, Akira KUREMATSU, Katsuhiko SHIRAI

Waseda University
Dept. of Computer Science and Engineering
Tokyo, Japan

Shigeki OKAWA

Chiba Institute of Technology
Narashino, Japan

ABSTRACT

The efficiency of multistream speech recognizers is investigated by performing several experiments. In order to take advantage of multistream features, each stream should compensate the weakness of the other streams. Our objective is to utilize frequency modulation (FM) which can compensate errors from traditional analysis methods. In order to achieve informational independence from other features based on the spectral/time envelope of signals, our features do not contain amplitude information, but contain temporal structure information of frequency modulation. Our method is evaluated by the continuous digit recognition of noisy speech. We confirmed that our AM-FM combination method is efficient for noisy speech recognition.

Index Terms— Speech recognition, feature extraction, multistream features, temporal analysis, frequency modulation

1. INTRODUCTION

The efficiency of multistream speech recognizers is investigated by performing several experiments [1]. In order to take advantages of multistream features, each stream should compensate the weakness of the other streams.

Most of the speech recognizers use the features derived by a spectral envelope as the primary cue for speech recognition. A family of alternative features that can compensate the weakness of spectral envelope derived from amplitude modulation (AM) are also utilized in many speech recognizers [2, 3]. In this paper, we introduce features derived from frequency modulation (FM); these features can compensate errors from AM analysis.

Features based on FM of speech have been investigated by employing several methods. For example, Wang *et al.* employed the average instantaneous frequencies of signals [4]. Paliwal *et al.* proposed a method based on spectral centroids that depends on FM of signals [5]. Dimitriadis *et al.* employed the average of instantaneous frequencies weighted by amplitudes [6]. However, all the previous methods for employing FM are based on stochastic quantities such as the average frequency and the modulation percentage of temporal segments. These methods eliminate the temporal structures in an instantaneous frequency. In order to use the temporal structures of the instantaneous frequency, we propose a more straightforward feature extraction method.

The possibility of using FM for speech recognition is also confirmed by the listening experiments performed by Yoshida *et al.* [7]. In these experiments, it is confirmed that the reconstructed signal that preserves the zero-crossing points of narrowband waveforms are perceivable through human speech recognition. Although most of the previous FM methods also use an amplitude factor in order to improve the performance [4–6, 8], the results of human speech recognition experiments indicate that the human speech recognizer can recognize speech only by using FM.

Our studies are motivated by this result. We consider that FM can be a cue for speech recognition by itself. The FM features derived without amplitude factor are considered independent of AM. Thus, the properties of the features are quite different from those of traditional features and are appropriate for multistream speech recognizers.

In this paper, in section 2, we first introduce our feature extraction frontend. The experimental setup is presented in section 3, and the results are discussed in section 4.

2. THE COMBINATION MODEL OF AM AND FM (AFMC)

Fig. 1 depicts a brief overview of our method, the AM-FM combination (AFMC).

Input signals are classified under monophones frame by frame by an AM classifier and an FM classifier. Subsequently, evidence obtained by both the classifiers are merged by an evidence merger component. Finally, the speech recognition results are determined by decoding the sequence of the merged evidence.

2.1. AM Classifier (HATS)

We employ the HATS method introduced by Chen *et al.* as an AM classification method [3]. This method can capture the amplitude modulation of speech signals efficiently.

In this section, we describe the HATS method.

2.1.1. AM Emphasis Using MLP-OL

Fig. 2 shows the block diagram of a HATS classifier.

First, input signals are separated by a Bark filterbank [9]. Subsequently, the output of the filterbank is processed by MLP-OL[1].

MLP-OLs are used to extract the significant modulation components from an envelope. MLP-OL is the general MLP classifier during the training phase (Fig. 3).

The input signal x_i of the i^{th} neuron in the input layer of the MLP-OL at the n^{th} frame is defined by

$$x_i = E_b\left(n + i - \frac{L+1}{2}\right). \qquad (1)$$

Here, L is the number of dimensions of the input vector (must be odd) and $E_b(n)$ is the energy of the output of the b^{th} channel of the

[1] MLP-OL stands for MLP minus output layer

filterbank at time n. Typically, the frame rate of E_b is set to 100 Hz, and L is set to 51.

As is typical for the MLPs trained to estimate the posterior probabilities, all the MLPs are trained using the teaching signal, which is "1.0" for the monophone associated with the central frame and "0" for all the others. We use the standard error back-propagation algorithm to optimize the weights of connections between layers so that the mean squared error is minimized.

During the application phase, the output layer of the MLP is removed. Because the input vector x_i can be interpreted as the time-series signal, the output of hidden neurons can be interpreted as the convolution of x and the weights between input neurons and the hidden neuron with a nonlinear sigmoid function. Therefore, the output of hidden neurons has a fixed frequency response that can improve the distinguishability of x. The filter constructed using the above procedure is called a "matched filter."

2.1.2. *Tandem Approach for Acoustic Modeling*

To recognize the output of matched filters, the HMM/MLP tandem approach is used in HATS [1]. In this approach, the input feature vector is classified under monophones by an MLP.

The MLPs in the tandem approach are also trained to estimate the posterior probabilities of the associated monophones; the teaching signal for training is "1.0" for the monophone associated with the central frame and "0" for all the others.

2.2. FM Classifier

In order to apply the advantages of HATS, which can find a matched modulation component from the training data, we apply the HATS method to an FM signal.

Fig. 4 shows the block diagram of an FM classifier.

2.2.1. *FM Extraction*

Several methods are proposed for AM-FM decomposition, such as the Teager energy operator (TEO) method [10] and the method based on the Hilbert transform [11]. Since our first motivation is based on the human perception of the zero-crossing points of signals, we define FM of speech signals by employing the zero-crossing points of the signals.

The logarithmic pseudo-instantaneous frequency (LPIF) is obtained by performing the following steps:

1. Measure the time interval $D(n)$ between the preceding and the following zero-crossing points for each sample.
2. The LPIF ($P(n)$) at time n is defined by $\log(\pi/D(n))$.

LPIFs can be considered as variants of the zero-crossings with peak amplitude (ZCPA) features [8] in which amplitude weighting is omitted. Amplitude weighting can improve the distinguishability of features. However, weighting makes features dependent on AM information. As our objective is to compensate the weakness of AM features, informational independence is important.

We take the average of the LPIF signal for each 25 ms window and then slide the window by 10 ms in order to achieve an equivalence between the frame rate of FM and AM features.

2.2.2. *FM Emphasis*

First, the input signal are separated by a Bark filterbank , which is used in the AM classifier. Because successive processes require time-domain signals, filters are implemented using FIR filters.

Subsequently, LPIF extraction is performed at each channel output in the filterbank.

The input signal x_i of the i^{th} neuron in the input layer of the MLP-OL at the n^{th} frame is defined by

$$x_i \quad = \quad P_b\left(n+i-\frac{L+1}{2}\right). \tag{2}$$

Similar to the AM classifier, we employ an MLP to classify the outputs of FM-matched filters under monophones.

2.3. Evidence Merger

Now, we have observed two streams of the MLP evidence. We merge them before using them as features of Gaussian mixture hidden Markov models.

We use the entropy-based combination of tandem acoustic models, which was introduced by Ikbal *et al.* [12].

First, we calculate the approximate posterior probability for the m^{th} MLP $p(c|x^m)$ with the expression

$$p(c|x^m) \quad = \quad \frac{(\exp(-y^m_{i(c)})-1)^{-1}}{\sum_{d\in C}(\exp(-y^m_{i(d)})-1)^{-1}}. \tag{3}$$

Here, x^m is the input vector of the m^{th} MLP; y^m, the output vector of the m^{th} MLP; C, the set of target classes (in this study, C

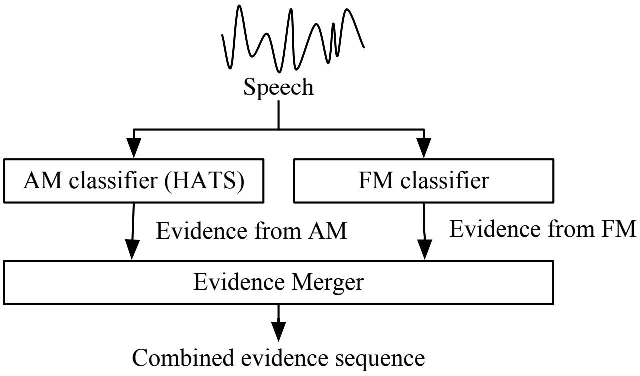

Fig. 1. Brief overview of proposed method.

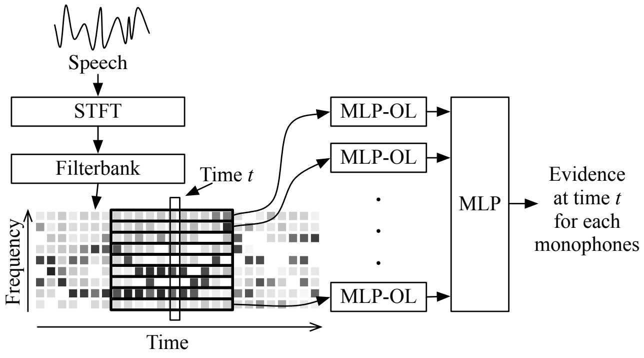

Fig. 2. Block diagram of HATS.

is the set of monophones); and $i(c)$, the mapping from the elements of C to the dimension index of y^m. This equation implies the cancellation of the sigmoid function and the application of the soft-max transfer function to the output of the MLP.

In concrete terms, x^1 is the AM feature vector; x^2, the FM feature vector; y^1, the output of the AM classifier; and y^2, the output of the FM classifier.

The conditional entropy of x^m is estimated as

$$H^m(C|x^m) \;=\; \sum_{c\in C} -p(c|x^m)\log p(c|x^m). \tag{4}$$

The weights of y^m are determined by taking the inverse of the conditional entropy h^m:

$$w^m \;=\; \frac{\{H^m(C|x^m)\}^{-1}}{\sum_{j=1}^{M}\{H^j(C|x^j)\}^{-1}}, \tag{5}$$

where M is the number of MLPs (in this study, M is 2).

Finally, we obtain the merged output $\hat{y}$ from the expression

$$\hat{y_i} \;=\; \sum_{m=1}^{M} w^m \log(y_i^m). \tag{6}$$

Because the dimensions of $\hat{y}$ correlate with each other, $\hat{y}$ is incompatible with the diagonal GMMs. It is necessary to transform $\hat{y}$ for dimensionality reduction and decorrelation. For this, we use the Karhunen-Loeve transformation (KLT).

3. EXPERIMENTS

In this section, we evaluate the performance of the AFMC by performing experiments. We conducted the continuous digit recognition of noisy speech in this experiment.

The training set and test set are taken from CENSREC-1 [13], which is the Japanese translation of the AURORA-2 data set. The training set used for both the MLP and HMM comprises 8,440 utterances of clean speech from 110 speakers. We select four noise environments from CENSREC-1 (restaurant, street, station, and airport) for the test. The test set comprises by 1,001 utterances for each noise environment and each signal-noise-ratio (SNR) condition.

The baseline comprises the MFCC and energy feature extraction system with cepstral mean normalization and it is augmented by the derivation and acceleration of the MFCC and energy. (MFCC_E_D_A_Z; 39 dims.)

The sample rate of speech signals in the experiments is fixed to 16,000 Hz. Therefore, the Bark filterbank splits them into 14 filtered signals.

All MLP-OLs are constructed with 20 hidden neurons, which means that we extract 20 modulation components for each filtered signals. Therefore, the number of features for MLP is the product of the number of bands (14) and 20. The number of hidden neurons for MLPs is fixed to 200.

We compare several methods, as described below.

- AIF
  This is a recognizer based on AIF features that are defined by removing the ALE features from the AIF/ALE [4]. (augmented by its derivations and accelerations; 42 dims.) AIF features are calculated by the DCT decorrelation of the segmental mean of the instantaneous frequency of narrowband signals.
- AM (HATS)
  A recognizer based on AM features is constructed by removing the FM classifier from the AFMC. This method is equivalent to HATS. The number of features is 280.
- FM
  A recognizer based on FM features is constructed by removing the AM classifier from the AFMC. The number of features is 280.
- AFMC
  This is the proposed method. The input feature vector size for each MLP is 280. The total number of input features is 560.

4. DISCUSSIONS

From Fig. 5, it is observed that the proposed AM-FM system exhibits considerable improvement as compared to other methods. In comparison with the MFCC feature extraction system proposed combination method reduced 43.6% of the word error at an SNR of 10 dB.

Although information on energy or amplitude is not included in the FM features, these features have achieved sufficient performance. It is considered that the temporal structures of LPIF contain phonetic information, although the information does not explicitly contains spectral/temporal envelope information. In comparison with AIF, our FM method exhibits considerable improvement. The effectiveness of the method, which captures the temporal structures of FM by using MLP-OLs, is confirmed by the experiment.

The results shows the indisputable fact that spectral envelope (MFCC) and time envelope (AM) are important for speech discrimination. However, the results also indicate that the FM of speech

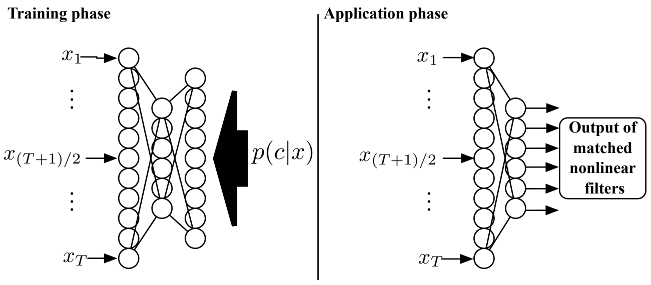

Fig. 3. Diagram of MLP-OL.

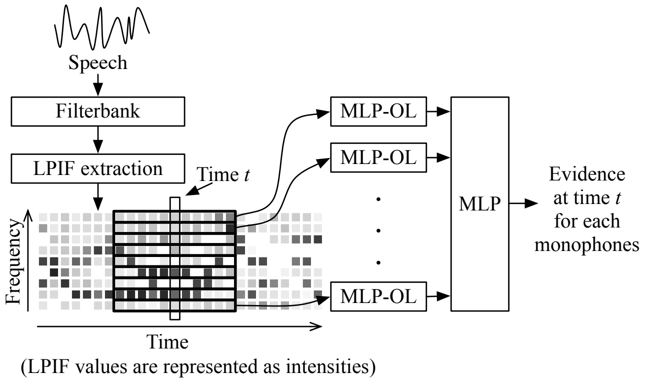
Speech
Filterbank
LPIF extraction
Time t
Frequency
Time
(LPIF values are represented as intensities)
MLP-OL
MLP-OL
MLP-OL
MLP
Evidence at time t for each monophones

Fig. 4. Block diagram of FM classifier.

Table 1. Average weights allocated on AM and FM classifiers and error reduction rate of AFMC comparing with AM.

	AM weight (%)	FM weight (%)	Err. Reduction Rate (%)
Clean	81.35	18.65	35.4
Restaurant (10 dB)	76.31	23.69	10.3
Street (10 dB)	61.76	38.24	26.9
Station (10 dB)	50.26	49.74	17.7
Airport (10 dB)	55.85	44.15	23.5

contains phonetic information and it can complement the AM features.

The relation between the allocated weights and the error reduction rate compared with AM is depicted in table 1. The table shows that noisy conditions damage the entropy of AM streams. Therefore, the relative availability of FM streams is increased. However, it appear that there is no linkage between the FM weight and the error reduction induced by the FM classifier. Although the combination method achieves higher accuracy, the weighting method based on inverse entropy might not be the optimal combination method.

5. CONCLUSIONS

In this paper, we introduced our novel feature extraction frontend (AFMC) that consists of an AM classifier, FM classifier, and evidence merger. Because the FM classifier can compensate the errors arising from the AM classifier, a combination of these classifiers is efficient for speech recognition.

We evaluated the proposed system by conducting the noisy digit recognition task. Our method reduced 43.6% of the word error at an SNR of 10 dB. The results show that our FM analysis method outperforms the FM analysis method employed in the previous AM-FM method. Therefore, it is confirmed that the time structures of FM are important for speech recognition.

Finally, we confirmed that the FM of speech can complement AM features.

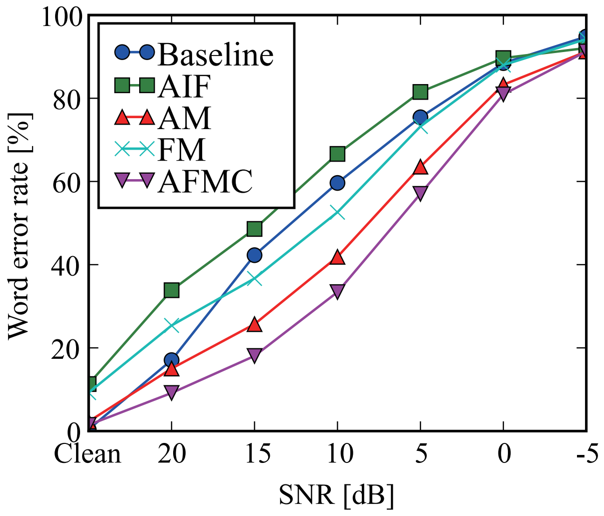

Fig. 5. Word error rate for noisy speech as a function of SNR.

6. ACKNOWLEDGEMENTS

This study was supported by the Advanced Research Institute for Science and Engineering, Waseda University, under the project "Research on Multi-Modal Human Interface Aiming for Spontaneous Communication System." It was partially supported by a Grant-in-Aid for Scientific Research (B) (17300066, 2007) from the Ministry of Education, Science, Sports and Culture, Japan.

7. REFERENCES

[1] N. Morgan, H. Bourlard, "An Introduction to he Hybrid HMM/Connectionist Approach," IEEE Signal Processing Magazine, pp. 25–42.

[2] H. Hermansky, N. Morgan, "RASTA Processing of Speech," IEEE Transactions on Speech and Audio Processing, Vol. 2, No. 4, pp. 578–589, October 1994.

[3] B. Chen, S. Chang, S. Sivadas, "Learning Discriminative Temporal Patterns in Speech: Development of Novel TRAPS-like Classifiers," Proc. Eurospeech, September 2003.

[4] Y. Wang, J. Hansen, G.K. Allu, R. Kumaresan, "Average Instantaneous Frequency (AIF) and Average Log-envelopes (ALE) for ASR with the Aurora 2 Database," Proc. Eurospeech, September 2003.

[5] J. Chen, Y. Huang, Q. Li, K.K. Paliwal, "Recognition of Noisy Speech Using Dynamic Spectral Subband Centroids," IEEE Signal Processing Letters, Vol. 11, No. 2, pp. 258–261, February 2004.

[6] D. Dimitriadis, P. Maragos, A. Potamianos, "Robust AM-FM Features for Speech Recognition," IEEE Signal Processing Letters, Vol. 12, No. 9, pp. 621–624, September 2005.

[7] K. Yoshida, M. Kazama, M. Tohyama, "Pitch and Speech-rate Conversion using Envelope Modulation Modeling," Proc. ICASSP-2002, Orland, I. 435–428.

[8] B. Gajić, K.K. Paliwal, "Robust Speech Recognition Using Features Based on Zero Crossings with Peak Amplitudes" Proc. ICASSP-2003, Hong Kong, I. 62–67.

[9] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," Journal of the Acoustical Society of America, Vol. 87, pp. 1738–1752, April 1990.

[10] J.F. Kaiser, "Some Useful Properties of Teager's Energy Operators," Proc. ICASSP-1993, Minneapolis.

[11] H. Suzuki, F. Ma, H. Izumi, O. Yamazaki, S. Okawa, K. Kido, "Instantaneous Frequencies of Signals Obtained by the Analytic Signal," Journal of Acoust. Sci. & Tech., Vol. 27, No. 3, 2006.

[12] S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, H. Bourlard, "Entropy Based Combination of Tandem Representations for Noise Robust ASR," Proc. INTERSPEECH-ICSLP-2004, Jeju Island, Korea, October 2004.

[13] CENSREC-1: http://sp.shinshu-u.ac.jp/CENSREC/ja/CENSREC/AURORA-2J/