A Comparative Study on AM and FM Features

Yotaro Kubo[1], *Shigeki Okawa*[2], *Akira Kurematsu*[1], *Katsuhiko Shirai*[1]

[1]School of Science and Engineering, Waseda University, Tokyo, Japan
{yotaro,shirai}@shirai.cs.waseda.ac.jp, a-kurematsu@aoni.waseda.jp,
[2]Chiba Institute of Technology, Narashino, Japan
okawa.shigeki@it-chiba.ac.jp

Abstract

In this paper, we investigate the advantages of frequency modulation (FM) features by conducting speech recognition experiments and statistical analysis. The importance of temporal aspects in speech recognition has been discussed along with the importance of amplitude modulation (AM) and frequency modulation. Recently, we have proposed a speech recognition system that is based on the combination of AM and FM features and confirmed its efficiency experimentally. Although the proposed speech recognizer assumes complementarity between the AM and FM features, it was not evaluated in previous studies. In this paper, in order to evaluate the complementarity between two types of features, we conducted continuous digit recognition tasks in artificial noisy conditions. We confirmed that the error rates of each classifier are significantly different depending to kind of noise. Then, we evaluated the statistical independency between these two types of features. We confirmed that the behaviors of these features are independent in realistic noisy environments.

Index Terms: Speech recognition, temporal features, robustness, frequency modulation

1. Introduction

The importance of temporal aspects in speech recognition has been discussed along with the importance of amplitude modulation (AM) and frequency modulation (FM). Features based on the FM of speech have been investigated by employing several methods. For example, Wang *et al.* employed the segmental average instantaneous frequencies of signals [1]. Paliwal *et al.* proposed a method based on spectral centroids that depends on FM of signals [2]. Dimitriadis *et al.* employed the average of instantaneous frequencies weighted by amplitudes [3].

The possibility of using temporal analysis of FM features for automatic speech recognition (ASR) are confirmed by hearing experiments conducted by Kazama *et al.* [4]. In these experiments, it is confirmed that the information on narrowband carrier signals contributes to the intelligibility of speech signals when the carriers are analysed using long-term analysis windows.

Motivated by the results of these hearing experiments, we proposed a temporal trajectory analysis of FM for ASR. We confirmed the efficiency of the combination of the AM and FM features in our recent experiments [5, 6].

Although the proposed speech recognizer assumes complementarity between the AM and FM features, it was not evaluated in previous studies. Therefore, the advantages of the FM analysis have not been clarified.

In this paper, we conduct several experiments to understand how FM features work in ASR. We show the complementarity between the AM classifier and the FM classifier by conducting speech recognition tasks in artificial noise conditions. Then, we show the statistical independency of these two types of features.

2. Classifiers for AM and FM

In this section, we present a speech recognizer, which is used in the speech recognition experiments in this paper.

2.1. AM Classifier (HATS)

We employ the HATS method introduced by Chen *et al.* as an AM classification method [7]. This method can efficiently capture the amplitude modulation of speech signals.

In this section, we describe the HATS method.

2.1.1. AM Emphasis Using MLP-OL

Fig. 1 shows the block diagram of a HATS classifier.

First, the input signals are separated by a Bark filterbank [8]. Subsequently, the output of the filterbank is processed by MLP-OL[1].

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

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

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

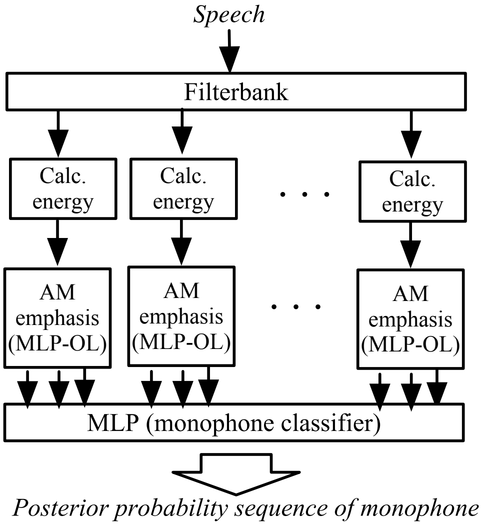

Figure 1: Block diagram of HATS.

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

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 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 backpropagation 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 [9]. 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. 2 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.

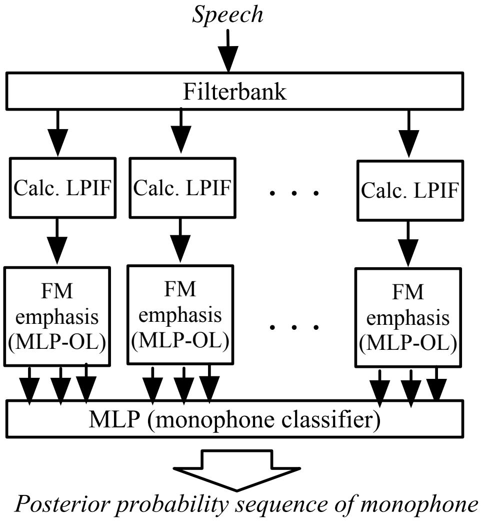

Figure 2: Block diagram of FM classifier.

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 [12] 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, input signals 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

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

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

3. Experiments and Discussions

We confirmed the efficiency of the AM/FM combination (AFMC) method by performing the noisy digit recognition task reported in [6]. The AFMC method reduced 43.6% of the word error at an SNR of 10 dB. The results also show that our FM classification method outperforms the FM analysis method employed in the previous AM-FM method.

In this paper, we focused on the reason behind the efficiency of the combination method. We discuss the results of the speech recognition experiments and statistical analysis.

3.1. Noisy Speech Recognition Experiments

In order to investigate the advantages of the FM analysis, we defined various artificial noises and evaluated the robustness for these noises.

The properties of noise that we considered are listed below.

- Stationary noise or burst noise,
- Narrowband noise or wideband noise (white noise).

We created different noise patterns by combining these properties.

- White noise (wn)
  stationary and full-range white noise.
- Band-pass filtered white noise (bpf_wn)
  The central frequency of a band-pass filter is obtained from uniform random values ranging from 1,000 Hz to 3,000 Hz, and the bandwidth is obtained from uniform random values ranging from 100 Hz to 2000 Hz.

- Burst noise (burst_wn)
  White noise of 250 ms duration and silence of 250 ms duration are added alternately

- Band-pass filtered burst noise (burst_bpf_wn)
  The parameters of a band-pass filter are same as bpf_wn.

The spectrograms of these noises are depicted in Fig. 3.

The training set is taken from CENSREC-1 [15], 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 added the 4 above-mentioned noises at 10 dB SNR to clean speech data in the test set of CENSREC-1, which comprised 1001 utterances from 104 speakers.

Table 1 shows the word accuracies of the AM and FM methods in the tested environments. From the results, we observed that the FM analysis has certain disadvantages with respect to narrowband noises. However, the FM analysis is advantageous for full-range noises. In contrast, the AM analysis is often degraded under full-range burst noise.

The error rates of each classifier are significantly different depending on characteristics of noise. Therefore, we confirmed that these two classifiers share a complementary relation.

3.2. Independency between AM and FM

Since all features are assumed to be dependent on phonetic information, most of features are interdependent in clean environments. However, under noisy conditions, some of the features are degraded by environmental noise. Therefore, since the corrupted features depend on the noise sources, they are independent of the robust features.

In order to understand the complementarity in a realistic noisy environment, we measured the independency of the features in noisy environments.

In order to support our argue on the independency of the behaviors of these features, a measure indicating the independency between non-stationary signals is necessary. We employed the concept proposed by Ando *et al.*, which can deal with non-stationary time-series variables [13]. According to this concept, the squared sum of segmental covariance indicates the independency of two variables.

Using this concept, we defined the independency measure $D(x, y)$ as follows:

$$D(x,y) = K(x,y) - \frac{K(x,x) - K(y,y)}{2}, \quad (3)$$

$$K(x,y) = \sum_t E_\tau[\tilde{x}(t,\tau) \cdot \tilde{y}(t,\tau)^T]^2 + C, \quad (4)$$

$$\tilde{x}(t,\tau) = x(t+\tau) - \overline{x}(t), \quad (5)$$

$$\overline{x}(t) = E_\tau[x(t+\tau)]. \quad (6)$$

Here, x denotes the original feature vectors, and $E[.]$ denotes the operations carried out to measure the empirical expectation by varying $\tau \in [-T, T](T = 25)$. The constant K ensures positive distances. In [13], it is suggested that T must be determined so that observed signals can be assumed to be quasi-stationary in $x_n(t + \tau)$.

Fig. 4 shows the component dependency matrix $D(x, y)$ of the features. The evaluated features are the energies (AM) and LPIFs (FM) of the output signals of the Bark filterbank (14 channels for AM features and FM features).

The analyzed signals are taken from multiconditional training set in CENSREC-1. Therefore, the independency is measured using speech signals that are corrupted by realistic environmental noise.

From the distance matrices, we confirmed that the AM features are highly dependent on each other. FM features are highly independent of all the other features, including the AM features. We confirmed that the AM features and FM features are independent of each other. Since the independency indicates that each type of features is robust to different types of noise, the complementarity between these two types of features is confirmed.

Several studies performed to understand the theoretical relation between AM and FM indicate that AM and FM are not completely independent [11, 14]. Our results show that the AM features are slightly dependent on the FM features in the neighbouring subband.

4. Conclusion

In this paper, we described the classifiers of AM patterns and FM patterns and discussed the complementarity of these classifiers.

We evaluated complementarity of each classifier by conducting continuous speech recognition experiments under artificial noisy conditions. We confirmed that the error rates of the two classifiers are significantly different depending on characteristics of noise.

Furthermore, the statistical independency between two types of features was measured. We confirmed that the AM and FM features are statistically independent. Therefore, we confirmed that these features have different characteristics when the signals are corrupted by environmental noise.

We confirmed that the AM classifiers and FM classifiers share a complementary relation. Therefore, the method that combines these classifiers functions effectively.

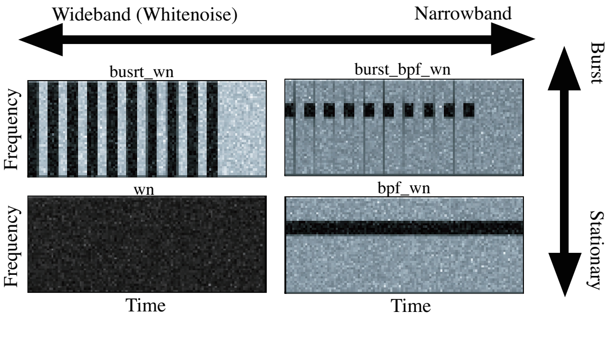

Figure 3: The spectrograms of selected noise patterns.

Table 1: Word accuracies in noisy environments (10 dB) of the compared methods as percentages.

	AM	FM
wn	45.53	49.46
bpf_wn	35.31	2.89
burst_wn	49.12	62.90
burst_bpf_wn	62.90	34.66

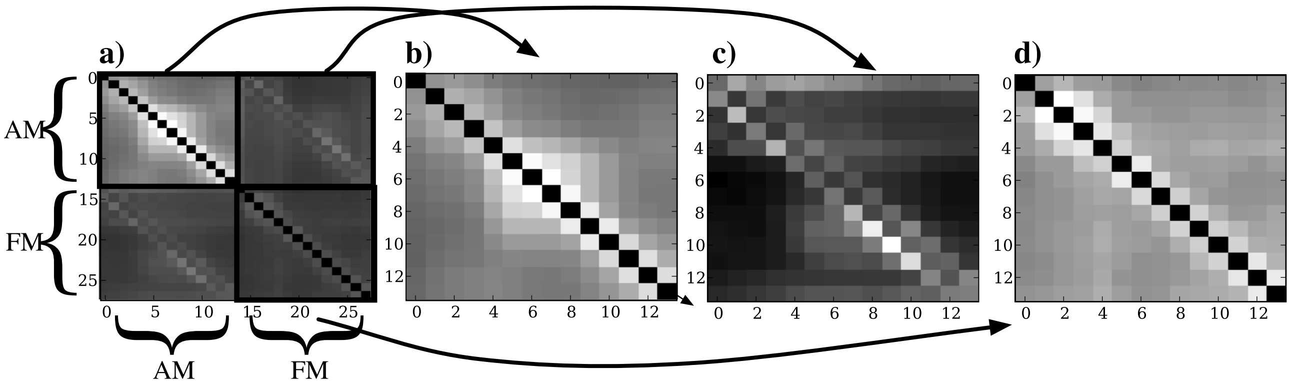

Figure 4: Component dependency matrices of the features. The dependency between the n^{th} component and the m^{th} component of the features are represented as intensity at point (m, n) (black: dependent, white: independent). (a) dependency between all features; (b) dependency matrix of AM features; (c) dependency between AM features and FM features; (d) dependency matrix of FM features.

5. 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 Hybrid Smart Room System."

6. References

[1] 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.

[2] 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.

[3] 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.

[4] M. Kazama, M. Tohyama, T. Houtgast, "Speech Reconstruction by Using Only Its Magnitude Spectrum Or Only Its Phase," Proc. 17^{th} International Congress on Acoustics, 2001.

[5] Y. Kubo, S. Okawa, A. Kurematsu, K. Shirai, "Recognizing Reverberant Speech Based on Amplitude and Frequency Modulation," IEICE Trans. on Inf. and Syst., VOL. E–61–D, No. 3, pp.448–456, March 2008.

[6] Y. Kubo, S. Okawa, A. Kurematsu, K. Shirai, "Noisy Speech Recognition Using Temporal AM-FM Combination," Proc. ICASSP-2008, Las Vegas, April 2008.

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

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

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

[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, pp. 163–170, 2006.

[12] 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.

[13] A. Ando, M. Iwaki, "Blind Separation of Nonstationary Sources by Block Decorrelation of Output Signal," Technical Report of IEICE (EA), Vol. 9, No. 57, September 2002.

[14] W. Nho, P.J. Loughlin, "When Is Instantaneous Frequency the Average Frequency at Each Time? ," IEEE Signal Processing Letters, Vol. 6, No. 4, pp. 78–80, April 1999.

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