A Study on Temporal Features Derived by Analytic Signal

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

Traditional feature extraction methods for automatic speech recognition (ASR), such as MFCC (Mel-frequency cepstral coefficients) and PLP (perceptual linear prediction) [6], are extracted from short-term spectral envelopes and can be used to realize promising ASR systems. On the other hand, features extracted by TRAPs-like classifiers [2] are based on long-term envelopes of narrow-band signals. These two forms of feature extractions use a mutual representation of energy in narrow band signals.

We have developed a feature extraction system that depends on not only the energy but also the modulation of carrier signals. Carrier signals involve attributes such as the spectral centroid, spectral gradient, number of zero-crossing points, and frequency modulation. Some experiments show that not only the spectral envelope and its modulation but also the zero-crossing points and frequency modulation form a significant portion of human speech perception [4].

In this study, we propose a method of carrier analysis, evaluate this method, and discuss the effectiveness of carrier analysis for ASR. Our method can reduce the phoneme error rate from 45.7% to 38.6%.

Index Terms: feature extraction, analytic signal, tandem approach, temporal feature

1. Introduction

The traditional features of automatic speech recognition (ASR), such as MFCC (Mel-frequency cepstral coefficients) and PLP (perceptual linear prediction) [6], are extracted from the short-term spectral envelopes of speech signals and can be used with hidden Markov models (HMMs) to realize promising ASR systems.

HMMs are based on the assumption that signals are stationary in specific segments; however, real speech signals are not stationary in nature. In order to capture the dynamic features of speech signals in their stationary form, it is conventional practice to augment these features by their derivatives and accelerations; however, this technique has its limitations.

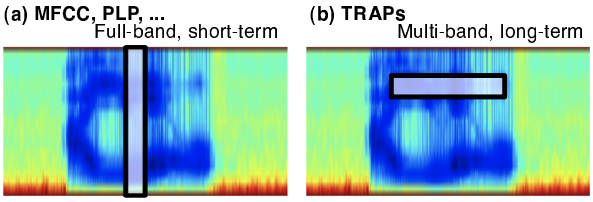

Figure 1: (a): Traditional feature; (b): TRAPS

These limitations motivate us to determine a more effective approach toward capturing dynamic features of speech signals. Hermansky *et al.* invented TRAPS (TempoRAl PatternS) to enable the effective use of the dynamics of narrow-band energy envelopes (Figure 1). The combination of TRAPS and PLP accomplishes high-accuracy speech recognition [1].

An advantage of dynamic features is that it alleviates unreliable cues from static features. Therefore, the incorporation of another feature derived by another attribute of the signals could improve the accuracy.

In general, a signal can be represented as a product of two signals; the envelope and the carrier. Traditional features and TRAPS are extracted using only envelopes. However, it was reported that carriers are also important for human speech perception [4]. Carrier signals involve attributes such as the spectral centroid, spectral gradient, number of zero-crossing points, and frequency modulation. These attributes are discarded in traditional feature extraction methods.

We focus on the effectiveness of carriers in speech recognition and propose a method to incorporate them.

In this paper, we first introduce our proposed system in sec. 2. The experimental results are provided in sec. 3, and these results are discussed in sec. 4.

2. S-LTHP System

Figure 2 shows the overview of our feature extraction system. We named our system "separated long-term Hilbert transform pair" (S-LTHP).

2.1. Preprocessing

First, the filter-bank that is designed to simulate frequency responses of human auditory perception is applied. Unlike TRAPS-like classifiers [2], this filter-bank is implemented using FIR filters.

The frequency response is defined by the following equations. (Figure 3)

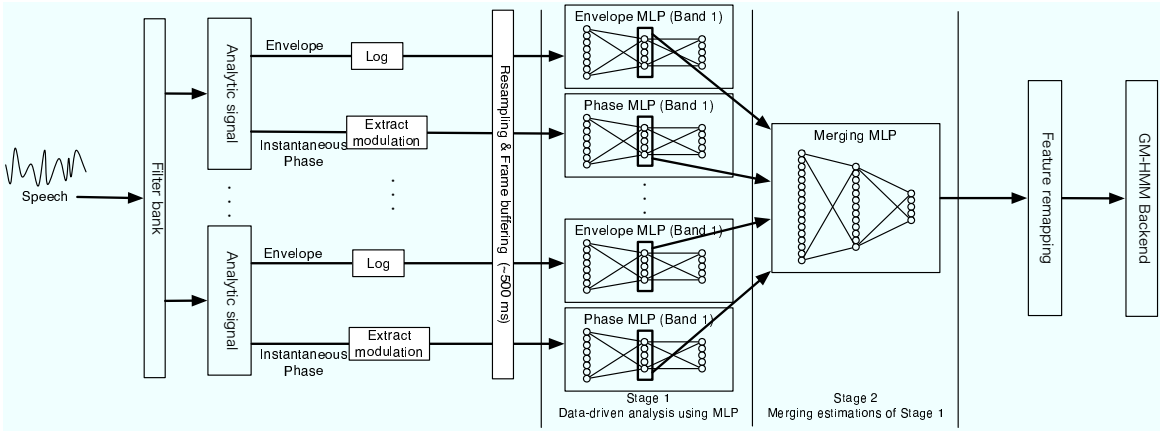

Figure 2: S-LTHP System Overview

$$\Omega(\omega) = 6\log\left[\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi}\right)^2 + 1}\right] \quad (1)$$

$$\Psi(\Omega) = \begin{cases} 0 & \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & -1.3 \le \Omega \le -0.5, \\ 1 & -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)} & 0.5 \le \Omega \le 2.5, \\ 0 & 2.5 < \Omega. \end{cases} \quad (2)$$

$$D_n(\omega) = \Psi(\Omega(\omega) - n) \quad (3)$$

Next, we separate these signals into envelopes and carriers. A narrow band signal $x(n)$ is generally expressed as

$$x(n) = \exp(m(n))\cos(\Omega_c n + \Phi(n)) \quad (4)$$

In this equation, Ω_c denotes the central frequency of the carrier signal and $\Phi(n)$ denotes the instantaneous phase modulation.

$\cos(\Omega_c n + \Phi(n))$ represents the carrier and depends on $\Phi(n)$. Therefore, $\Phi(n)$ is convenient for representing carrier signals.

$\exp(m(n))$ represents the envelope. The logarithmic envelope $m(n)$ is often used in other features such as MFCC, PLP, and TRAPS. In this model, the sign of $x(n)$ is depends on the carriers; therefore, $\cos(\Omega_c n + \Phi(n))$ involves zero-crossing points and spectral centroid.

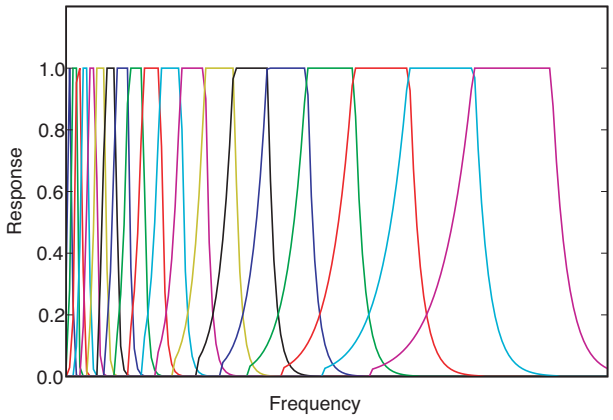

Figure 3: Frequency response of filter $D(n)$

In order to extract $m(n)$ and $\Phi(n)$ from $x(n)$, an analytic signal $x^+(n)$ is used.

The analytic signal is defined as follows:

$$x^+(n) = x(n) + j\hat{x}(n) \quad (5)$$

$$= \exp(m(n))\cos(\Omega_c n + \Phi(n)) + j\exp(m(n))\sin(\Omega_c n + \Phi(n)) \quad (6)$$

where $\hat{x}(n)$ is the Hilbert transform of $x(n)$.

The amplitude envelope $a(n)$ and instantaneous phase $\phi(n)$ are defined by the following equations;

$$a(n) = |x^+(n)| \quad (7)$$

$$\phi(n) = \arg x^+(n) \quad (8)$$

Next, we calculate $\Phi(n)$ and $m(n)$ by using the following equations

$$m(n) = \log(a(n)) \quad (9)$$

$$\Phi(n) = \hat{\phi}(n) - \Omega_c n \quad (10)$$

$\hat{\phi}(n)$ denotes the unwrapped $\phi(n)$.

Ω_c is defined as

$$\Omega_c = \frac{\phi(N-1) - \phi(0)}{N} \quad (11)$$

where N is the length of signal $x(n)$.

Finally, we obtained two representations of the sub-band signal: $m(n)$ and $\Phi(n)$. We resampled these representations at 100 Hz and normalized the mean to 0.5 and the variance to 0.25 over each utterance. Figure 4 depicts $m(n)$ and $\Phi(n)$ of the fifth band of central phoneme class /r/.

2.2. MLPs

The input vector of the envelope MLPs at time t is $m(n)$ ($n \in [t-25, t+25]$). The input vector of the phase MLPs at time t is $\Phi(n)$ ($n \in [t-25, t+25]$).

As is typical for the MLPs trained to estimate posterior probabilities, all the MLPs are trained using output targets that

are "1.0" for the monophone associated with the central frame and "0" for all others. We have employed the standard error back-propagation algorithm as the training method.

Next, merging MLPs are used to merge the estimations of the sub-band MLPs. The input vector of the merging MLPs is defined by the concatenated vector of the output of the hidden layers in sub-band MLPs. The teaching signals are the same as the sub-band MLPs.

2.3. Feature Remapping

The output vector of the MLPs that approximate the posterior probabilities of the phonetic classes have a skew symmetry and are incompatible with the Gaussian mixture model (GMM).

It is necessary to adapt the feature distribution to the Gaussian model in order to achieve stable recognition.

We use logarithmic function for non-linearity and Karhunen-Loeve Transforming for orthogonalization.

3. Experiments

3.1. Basic Comparison

In this section, we evaluate the performance of the proposed system.

To ignore the performance of language models, we conducted context-free continuous phoneme recognition in this experiment. The training set used for both MLP and HMM training comprised approximately 16 h of spontaneous speech data from "the Corpus of Spontaneous Japanese" (CSJ). All these models were trained to be speaker and gender independent. The accuracy is measured using spontaneous speech data (approximately 1 h) from CSJ.

The baseline is the MFCC feature extraction system that is augmented by the derivations and accelerations of the MFCC and the derivation and acceleration of energy. (MFCC_E_D_A_N_Z, 38 dims.)

We selected a TRAPS-like classifier, called hidden-activation TRAPS (HATS), for comparison. Currently, HATS has achieved the highest accuracy among TRAPS-like classifiers [3].

Three variants of the HATS system are compared:

- HATS (26 dims.),
- MFCC (12 dims.) augmented by HATS (26 dims.; total: 38 dims.), and
- baseline (38 dims.) augmented by HATS (26 dims.; total: 64 dims.).

Three variants of the S-LTHP system are compared:

- S-LTHP (26 dims.),

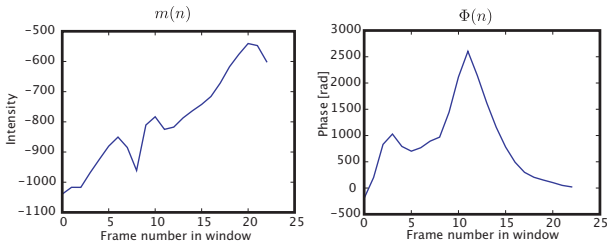

Figure 4: Left: $m(n)$; Right: $\Phi(n)$

Table 1: Phoneme error rate

System Description (# of dims.)	Phoneme Error Rate (%)	Error Reduction (% Rel.)
Baseline: MFCC_E_D_A_N_Z (38)	45.7	-
HATS (26)	47.3	-3.5
S-LTHP (26)	45.8	-0.2
HATS + MFCC (38)	43.1	5.6
S-LTHP + MFCC (38)	40.9	10.5
HATS + MFCC_E_D_A_N_Z (64)	39.4	13.8
S-LTHP + MFCC_E_D_A_N_Z (64)	38.6	15.5

Table 2: Frame error rate of variants

System Description	Frame Error Rate (%)	Error Reduction (% Rel.)
Baseline: HATS	25.1	-
IPP	41.3	-64.5
LTHP	24.1	4.0
S-LTHP	22.1	12.0

- MFCC (12 dims.) augmented by S-LTHP (26 dims.; total: 38 dims.), and
- baseline (38 dims.) augmented by S-LTHP (26 dims.; total: 64 dims.).

Table 1 summarizes the phoneme error rate ($\%insertion + \%deletion + \%substitution$) of the systems being compared.

3.2. Comparing Variation Systems

In order to evaluate the effectiveness of instantaneous phases, we conducted another comparison experiment. Many factors influence the phoneme error rate in experiments that use HMMs. Therefore, we examined the system using the frame accuracy of merging MLP's output vector.

We selected two variants of S-LTHP for comparison. The first variant to be compared is instantaneous phase patterns (IPP; Figure 5-b). The HATS system is equivalent to S-LTHP system without phase MLPs. (Figure 5-a). We can consider the opposite classifier, an S-LTHP system without envelope MLPs. The accuracy of this model might indicate the effectiveness of instantaneous phase clearly.

We can consider several methods to combine the instantaneous phase and envelope. We examined another structure of combination that analyzes the instantaneous phases and envelopes simultaneously, called LTHP (Figure 5-c). The S-LTHP estimation process involves separation by the type of input signals; envelopes or instantaneous phases. In the concept of multi-band ASR, it is considered that separation by sub-band is effective. However, the effectiveness of separation by the type of input signals is unclear.

All MLPs (excluding merging MLP) used by HATS, IPP, LTHP, and S-LTHP have the same number of neurons in the

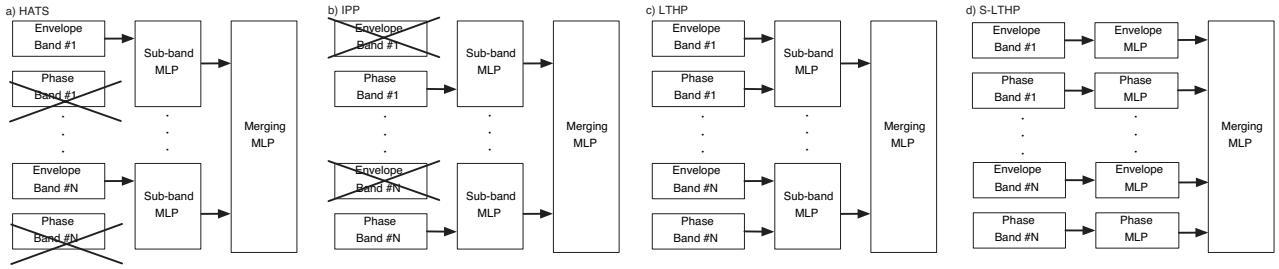

Figure 5: (a): HATS; (b): IPP; (c): LTHP; (d): S-LTHP

hidden layer.

Table 2 summarizes the frame error rate of the systems being compared. The frame error rate is measured by labeled spontaneous speech data (approximately 1 h) from CSJ that is not present in the training data set.

4. Discussions

From Table 1, it is observed that the proposed system exhibits considerable improvement from baseline and sufficient improvement from HATS. The best performance was achieved by the S-LTHP + MFCC_E_D_A_N_Z feature extraction method, which showed a gain of 38.6% in the phoneme error rate and reduced 15.5% of the phoneme error from the baseline system (MFCC_E_D_A_N_Z). In the comparison with HATS, our system reduced approximately 3% of the phoneme error.

Table 2 clearly shows the effectiveness of the instantaneous phase analysis. We observed that IPP models have poor accuracy (41.3% in the frame error rate); however, they might be suitable for ASR. The input of the IPP models is the resampled modulation component of the xinstantaneous phase. Therefore, this component contains phonetic information which is less than that in the resampled envelopes.

Conventional studies reported that the resampling of envelopes rarely results in a loss of phonetic information. However, we confirmed by preliminary listening experiments that the resampling of instantaneous phases often results in such a loss. Another method for carrier analysis without resampling might improve the accuracy.

The results of LTHP and S-LTHP models indicate that the structural constraints implemented by S-LTHP can eschew the local optimum in their training and utilize neurons in hidden layers more efficiently. Therefore, it is preferred that instantaneous phases and envelopes are discriminated separately.

5. Conclusions

In this study, we proposed a method to analyze carriers of narrow-band signals using a tandem approach, called S-LTHP. S-LTHP is the method that analyzes temporal patterns of analytic signal using MLPs.

We evaluated the performance of the proposed method by continuous phoneme recognition without using any language models. The proposed system exhibits considerable improvement from baseline and sufficient improvement from HATS.

To confirm the effectiveness of the carrier analysis, we have developed two variants of S-LTHP: IPP and LTHP and evaluated the output of the MLPs and their frame error rates. We confirmed that the incorporation of long-term carrier analysis is effective for improving the accuracy of speech recognition, and the structural constraints implemented by S-LTHP can eschew the local optimum in their training.

6. Acknowledgements

This study was supported by the Advanced Research Institute for Science and Engineering of Waseda University under the project "Research on Multi-Modal Human Interface Aiming for Spontaneous Communication System" and was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 17300066, 2005.

7. References

[1] H. Hermansky, S. Sharma, "TRAPS - Classifiers of temporal patterns," Proc. ICSLP '98, Sydney, Australia, November 1998.

[2] B. Chen, S. Chang, S. Sivadas , "Learning long term temporal features in LVCSR using neural networks," Proc. ICSLP '2004, pp. 612–615.

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

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

[5] H. Hermansky, "Should recognizers have ears?" Speech Communication, Invited paper, 25 (1–3):3–27, 1998.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Journal of the Acoustical Society of America, vol. 87, pp. 1738–1752, Apr. 1990.

[7] N. Morgan, Q. Zhu, A. Stolcke, K. Sönmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, Ö. Çetin, H. Bourlard, M. Athineos "Pushing the Envelope – Aside," IEEE Signal Processing Magazine, September 2005.

[8] 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. of the Eurospeech, 2003

[9] H. Hermansky, D.P.W. Ellis, S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," Proc. ICASSP-2000, Istanbul.

[10] J. Bilmes, "Maximum mutual information based reduction strategies for cross correlation based joint distributional modeling," Proc. ICASSP-98, Seattle, pp. 469–472.