Intra- and Inter-frame Features for Automatic Speech Recognition

Sung Joo Lee, Byung Ok Kang, Hoon Chung, and Yunkeun Lee

In this paper, alternative dynamic features for speech recognition are proposed. The goal of this work is to improve speech recognition accuracy by deriving the representation of distinctive dynamic characteristics from a speech spectrum. This work was inspired by two temporal dynamics of a speech signal. One is the highly non-stationary nature of speech, and the other is the inter-frame change of a speech spectrum. We adopt the use of a sub-frame spectrum analyzer to capture very rapid spectral changes within a speech analysis frame. In addition, we attempt to measure spectral fluctuations of a more complex manner as opposed to traditional dynamic features such as delta or double-delta. To evaluate the proposed features, speech recognition tests over smartphone environments were conducted. The experimental results show that the feature streams simply combined with the proposed features are effective for an improvement in the recognition accuracy of a hidden Markov model–based speech recognizer.

Keywords: Speech recognition, feature extraction.

I. Introduction

The recent advances of speech recognition technology and the widespread use of smartphones have made it possible to search the internet using voice technology. In addition, people have begun talking to their smartphones to send messages, make phone calls, set reminders, and more. These trends in smartphone use have made it easy to obtain massive amounts of speech data from ordinary users, and these voice footprints have fertilized modern speech-recognition technology. However, the recognition accuracy of even state-of-the-art speech recognizers still remains insufficient for smartphone users. Among the component technologies composing a modern speech-recognition system, feature extraction is one of the most important steps with the potential for a recognition performance breakthrough. Despite extensive research efforts in feature extraction, perfect features have yet to be introduced. Unfortunately, feature extraction still remains a challenging research field. In this work, we focus on dynamic feature extraction to improve the accuracy of speech recognition. This work is motivated by two temporal dynamic characteristics of speech; one being the highly non-stationary nature of speech. Since the spectral transition sometimes changes so rapidly, it neutralizes the short-time quasi-stationary assumption of speech signal; this property of speech has yet to be investigated. The other characteristic is the inter-frame change of the speech spectrum. The idea of using inter-frame features has already been introduced in traditional feature extraction methods—namely, the “delta” and “double-delta” [1]. Although massive speech data can be achieved these days, the recognition performance of the traditional dynamic features is limited since they describe only the simplified temporal dynamics (increase or decrease, acceleration or de-acceleration) of speech. In this work, we think that speech recognition performance can be improved by exploiting the more detailed temporal variation of a spectral envelope. To do so, we adopt a temporal discrete cosine transform (DCT) method.

II. Conventional Feature Extraction Methods

The aim of the feature extraction is to compute feature vectors that provide a compact representation of phoneme

Manuscript received Apr. 16, 2013; revised Sept. 10, 2013; accepted Nov. 5, 2013.

This work was supported by the IT R&D Program of MSIP/KEIT (10035252, Development of Dialog-Based Spontaneous Speech Interface Technology on Mobile Platform).

Sung Joo Lee (phone: +82 42 860 5732, lee1862@etri.re.kr), Byung Ok Kang (bokang@etri.re.kr), Hoon Chung (hchung@etri.re.kr), and Yunkeun Lee (yklee@etri.re.kr) are with the SW · Content Research Laboratory, ETRI, Daejeon and also with the Department of Broadband Network Technology, University of Science and Technology, Daejeon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.14.0213.0181

identification from an input speech signal, while suppressing intrinsic variations of speech, such as gender, age, dialect, speaking rate, and so on. The traditional feature extraction methods can be categorized into several groups according to their signal analysis techniques as follows:

1. Spectral analysis–based approaches: MFCC [2].
2. LPC analysis–based approaches: PLP [3].
3. Time-frequency analysis–based approaches: GTCC [4].
4. Human auditory model–based approaches: ZCPA [5].

Among the traditional methods, the Mel-frequency cepstral coefficient (MFCC) is widely used and known as a robust feature, while its algorithm is relatively simple. However, its recognition performance is not sufficient for ordinary users. Perceptual linear prediction (PLP) cepstrum has the potential for more precise spectral-envelope estimation, and its recognition performance is competitive with MFCC. Therefore, PLP is also a popular feature for voice recognition. Zero-crossings with peak amplitude method is known as a robust feature in the presence of background noise, though its phoneme identification accuracy is relatively lower than other spectral envelope–based approaches. Recently, researchers studying a speech front-end are interested in a gamma-tone cepstral coefficient (GTCC) owing to its positive contributions to recognition accuracy. Traditional feature extraction methods employ various frameworks and concepts to extract proper features for voice recognition; there are two common ideas: one is the short-time quasi-stationary assumption of a speech signal, and the other is to exploit the temporal changes of a speech spectrum using dynamic features; that is, portions of the input signal are separated into sequential frames with a predefined size (20 ms to 25 ms) before being analyzed. Static features are derived in a frame-by-frame manner. The dynamic features delta and double-delta are then extracted from sequential static features. The aim of the dynamic features is to alleviate the temporal independency of the hidden Markov model–based speech recognizer. However, since the traditional dynamic features describe the simple temporal variation of speech, more detailed dynamic features are necessary to improve recognition accuracy.

III. Proposed Intra- and Inter-frame Features

This work is motivated by two different temporal dynamic characteristics of speech. One is intra-frame variations, caused by the highly non-stationary nature of speech. The other is the inter-frame changes of the speech spectrum. Both are caused by continuous articulatory movements. In this work, we attempt to derive these speech characteristics from an input signal in a simple manner and investigate their feasibility as additional features for speech recognition. The intra-frame variations are captured through a simple sub-frame analyzer, and the inter-frame variations are measured by the temporal DCT. Unlike traditional dynamic features (delta and double-delta), we try to estimate temporal fluctuations of a more complex nature from a speech spectrum. Our feature extraction method is based on a gamma-tone filter bank since the time-frequency signal analysis approach is good for a sub-frame analysis. Otherwise, additional fast Fourier transform (FFT) steps are required for sub-frame analysis in the case of a spectral analysis–based feature extraction approach such as MFCC.

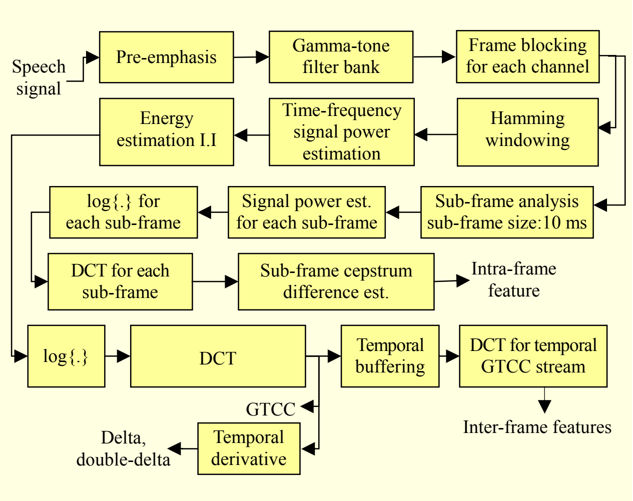
Speech signal
Pre-emphasis
Gamma-tone filter bank
Frame blocking for each channel
Energy estimation I.I
Time-frequency signal power estimation
Hamming windowing
log{.} for each sub-frame
Signal power est. for each sub-frame
Sub-frame analysis sub-frame size:10 ms
DCT for each sub-frame
Sub-frame cepstrum difference est.
Intra-frame feature
log{.}
DCT
Temporal buffering
DCT for temporal GTCC stream
GTCC
Inter-frame features
Delta, double-delta
Temporal derivative

Fig. 1. Block diagram of proposed GTCC.

Figure 1 describes the block diagram of our proposed intra- and inter-frame feature extraction procedure. Unlike the original GTCC, we adopt a simplified pre-emphasis filter instead of the equal loudness pre-emphasis technique, and the overall procedures become closer to the MFCC. As shown in Fig. 1, the highly non-stationary nature of speech is captured by using a sub-frame analyzer. The spectral variation within a speech analysis frame is represented by sub-frame cepstral differences. For the intra-frame feature extraction, an analysis frame (20 ms) is divided into two sub-frames, each 10 ms in size. After the completion of gamma-tone filter banking [4], the *m*th filter-bank output signal energy for each sub-frame is calculated as follows:

$$E_{m,1} = \sqrt{\sum_{n=0}^{(N/2)-1} \left[X_m(n)W(n)\right]^2},$$

$$E_{m,2} = \sqrt{\sum_{n=(N/2)}^{N-1} \left[X_m(n)W\left(n-{}^{N}\!/_{2}\right)\right]^2}, \qquad (1)$$

where $X_m(n)$ and $W(n)$ indicate *m*th filter-bank output signal and window function, respectively. Frame size is denoted by *N*. The logarithmic representations for each sub-frame are obtained as follows:

$$LogE_{m,1} = \log\lfloor E_{m,1} \rfloor,$$

$$LogE_{m,2} = \log\lfloor E_{m,2} \rfloor. \qquad (2)$$

The sub-frame cepstral differences are obtained as follows:

$$\begin{aligned} CD_k = &-\sum_{m=0}^{M-1} LogE_{m,1} \cos\left[\frac{\pi}{M}\left(m+\frac{1}{2}\right)k\right] \\ &+\sum_{m=0}^{M-1} LogE_{m,2} \cos\left[\frac{\pi}{M}\left(m+\frac{1}{2}\right)k\right], \end{aligned} \qquad (3)$$

where k and M indicate quefrency index and filter-bank number, respectively. The use of inter-frame features in speech recognition was originated by Furui, in 1986 [1]. The so-called "delta" and "double-delta" features are widely used owing to their positive contribution to speech recognition, and most modern speech-recognition systems conventionally append the dynamic features to the static features. However, the phoneme discriminability of the traditional dynamic features is limited since they merely represent simple temporal variations of speech envelope. In this work, we attempt to extract more detailed temporal spectrum fluctuations. We believe that one of the simplest methods for deriving dynamic components from a temporal input sequence is DCT. For example, an up-and-down spectral change can be measured using the first frequency component of DCT. Our previous experiment showed that the physical properties of the first and second frequency components are similar to the traditional dynamic features. In addition, various frequency components that represent the temporal variations of a speech spectrum can be derived from the temporal DCT. However, we are interested in lower-frequency components other than the DC term, since we assume that low-frequency components are more strongly related with the temporal dynamic nature of speech. After buffering several sequential static features, the temporal DCT coefficients are obtained as follows:

$$X_{m,k} = \sum_{n=0}^{T-1} x_{m,n} \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right], \qquad (4)$$

where m and k indicate quefrency and DCT index, respectively. The mth cepstral component at time n is denoted by $x_{m,n}$. In addition, T is temporal-frame buffer size. In this work, we use nine sequential frames for inter-frame feature extraction. This frame window size is related with the temporal-frame window size of the traditional methods when two left-and-right frames are considered.

IV. Experimental Results

The performance evaluation is done by comparing the speech-recognition rates with the traditional MFCC and PLP. The cepstral mean subtraction technique is applied to all the static features to normalize channel variation. The voice recognition task is a smartphone voice internet search in Korean. That is, the target evaluation task is large-vocabulary continuous speech recognition. The unique dictionary size is about 389 entries and the lexicon size is extended to 1,179,028 entries by considering variants in pronunciation. The language model (LM) was also built to cover various user requirements for voice internet service. The LM is estimated using huge internet search queries provided by Daum Communications (a commercial Web portal company in South Korea, http://www.daum.net). The speech corpus for acoustic modeling is composed of 1.6 million utterances (approx. 1,150 hours), including real data. A tri-phone–based acoustic model was built using the Hidden Markov model toolkit [6]. We collected the test corpus from ten ordinary users (five females and five males) using iPhones for a recognition result analysis. The speech data (3,000 utterances) are categorized into five groups depending on the speech acquisition environments (office, bus, restaurant, home and subway). Each data group is composed of 600 utterances. Several sound pressure levels are considered in this data collection to obtain more realistic speech data: normal speech, weak sound speech (such as whispering), and loud speech. Therefore, the speech-recognition rate, even in an office environment, is not very high. Test data other than from the office group are collected in the presence of typical background noises representative of the categories. At home, a television is activated as a noise source. All noise levels other than in an office environment are quite high, and the speaking styles are located somewhere between dictation and fluent speech. Therefore, pronunciation variants (such as weakness, omission, and co-articulation) often occur. Consequently, the overall speech-recognition accuracy of the baseline system is poor. Depending on how the proposed dynamic features are appended to the static features, various feature combinations are possible. In this work, we investigate several feature combinations as follows:

1. GTCC (13) + delta (13) + double-delta (13) + intra-frame features (13): vector dimension, 52.
2. GTCC (13) + inter-frame feature (3 DCT components except C0, 39): vector dimension, 52.
3. GTCC (13) + inter-frame feature (2 DCT components except C0, 26) + intra-frame feature (13): vector dimension, 52.
4. GTCC (13) + inter-frame feature (3 DCT components except C0, 39) + intra-frame feature (13): vector dimension, 65.

The purpose of the first feature combination is to evaluate the phoneme discrimination saliency of the proposed intra-frame

Table 1. Sentence recognition rate (%) under original data.

	Office	Bus	Restaurant	Home	Subway	Total
MFCC (39)	61.33	12.33	18.17	32.22	12.37	27.28
PLP (39)	59.67	15.33	22.50	30.72	14.63	28.57
GTCC (39)	60.67	16.83	21.50	34.22	13.76	29.40
Method 1	66.67	19.17	24.67	38.56	17.60	33.33
Method 2	67.83	23.67	27.33	40.07	20.03	35.79
Method 3	70.33	23.67	29.83	39.40	21.78	37.00
Method 4	71.83	26.67	29.67	42.74	22.82	38.75

Table 2. Sentence recognition rate (%) under enhanced data.

	Office	Bus	Restaurant	Home	Subway	Total
MFCC (39)	68.33	27.50	35.67	35.56	26.31	38.67
PLP (39)	67.33	31.50	40.50	35.56	30.31	41.04
GTCC (39)	69.08	32.33	37.83	38.23	28.57	41.21
Method 1	71.83	36.33	42.67	43.91	32.58	45.46
Method 2	73.67	37.00	46.00	45.74	35.37	47.56
Method 3	77.33	38.50	48.50	43.74	35.71	48.76
Method 4	76.50	41.83	48.83	46.91	36.06	50.03

feature. Since the third components of the proposed inter-frame features represent relatively fast temporal variation of speech envelope, the phoneme discriminability needs to be compared with the intra-frame feature. Therefore, the second and third feature combinations are prepared. The last feature combination is for investigating the overall performance of the proposed features. The similar idea of using temporal DCT–based augmented features is cited in [7]. However, the reported recognition performance was different from ours since the cepstral mean normalization technique was not applied in this past work. The speech-recognition experiment results under the original speech data are depicted in Table 1.

As shown in Table 1, we think that the proposed intra-frame feature set contains relatively good phoneme discriminability. Since most modern speech-recognition systems employ speech enhancement technology for environmental robustness, we also evaluate the speech-recognition rate in the presence of speech enhancement technology using a Wiener filter [8]. The speech-recognition results over the enhanced speech data are described in Table 2.

As shown in Table 2, all recognition rates over the enhanced speech data are better. It is also shown that the proposed dynamic features outperform traditional feature extraction methods in all cases. In addition, it seems that the proposed intra- and inter-frame features do not interfere with each other. We think that this is a desirable independency as an additional feature for speech recognition.

V. Conclusion

This work is a feasibility study on alternative dynamic features for speech recognition. We investigated the recognition performance improvement using the dynamic features that represent the dynamic nature of speech in more detail. We focused on two kinds of time-varying characteristics of speech. One is the intra-frame features that represent the highly non-stationary nature of speech, and the other is the inter-frame features that describe the temporal dynamics of speech in more detail. We appended the proposed dynamic features to the static GTCC features in several ways. The speech-recognition test results indicated that the proposed dynamic features were effective in improving recognition accuracy. It was also shown that the proposed intra- and inter-frame features did not interfere with each other in terms of recognition accuracy. We believe that this is a desirable independent characteristic as an additional feature for speech recognition.

References

[1] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 34, no. 1, Feb. 1986, pp. 52–59.

[2] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, Aug. 1980, pp. 357–366.

[3] H. Hermansky, "Perceptual Linear Prediction (PLP) Analysis of Speech," *J. Acoust. Soc. America*, vol. 87, no. 4, Apr. 1990, pp. 1738–1752.

[4] W.H. Abdulla, "Auditory Based Feature Vectors for Speech Recognition Systems," *Advances in Communications And Software Technologies*, WSEAS ed., Athens, Greece: WSEAS Press, 2002, pp. 231–236.

[5] D.-S. Kim, S.-Y. Lee, and R.M. Kil, "Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, Jan. 1999, pp. 55–69.

[6] S. Young et al., *The HTK Book (for HTK version 3.4)*, Cambridge, England: Cambridge University Engineering Department, 2006.

[7] B. Milner, "A Comparison of Front-End Configurations for Robust Speech Recognition," *Proc. ICASSP*, Orlando, FL, USA, vol. 1, May 13–17, 2002, pp. 797–800.

[8] S.J. Lee et al., "Statistical Model-Based Noise Reduction Approach for Car Interior Applications to Speech Recognition," *ETRI J.*, vol. 32, no. 5, Oct. 2010, pp. 801–809.