Temporal Information in Tone Recognition

Payton Lin[1], *Syu-Siang Wang*[1], *Yu Tsao*[1]

[1] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

ABSTRACT

Traditionally, only five components are regarded as having to do with the special characteristics of recognizing tones, while front-end processing and feature extraction have been considered essentially independent. Since mismatch between training and testing in signal-space leads to subsequent distortions in feature-space and model-space, determining whether front-end processing and feature extraction is independent or dependent will be critical for robustness.

Index Terms— Tone recognition, temporal features

1. INTRODUCTION

Nearly perfect recognition was computed under conditions of greatly reduced spectral information [1]. Spectrally-reduced data was synthesized by extracting the temporal envelopes of clean test data and modulating sinusoids generated at the center-frequency of each analysis band. Even though models were trained on the clean training database, 16-subbands of temporal envelope information was sufficient to achieve the accuracy attained with the original clean test data. When assessed in both the training and testing environments, the framework was shown to lower Kullback-Leibler divergence [2]. When expanded to tasks involving larger databases [3], performance depended on the periodic temporal structure of the signal carrier type. For instance, masking sidebands with white-noise carrier spectra [4] limited the delicate structure of attributes and showed improved performance when using triangular filter characteristics during feature extraction.

The original correspondence [1] reported that changing the bandwidth of the subband envelopes had no significant effect on a balanced database [5]. Since periodicity fluctuates at a rate which is directly related to the change in fundamental frequency (F0), the present study expands by investigating a tonal database distinguished by F0 contours [6]. Two low-pass filters with cutoff frequencies of 50 and 500 Hz will be used for envelope extraction in order to evaluate the relative contribution of temporal envelope (2-50 Hz) and periodicity (50-500 Hz). These two cutoff frequencies were selected to differentiate fine structure features (500-10,000 Hz) as in [7].

2. TEMPORAL FEATURE EXTRACTION

Input materials were first processed through a pre-emphasis filter (first-order Butterworth high-pass filter at 1200 Hz), and then band-pass filtered into N = [1, 2, 3, 4] subbands (fourth-order Butterworth filters) between 100 and 4,000 Hz. The equivalent rectangular bandwidth scale was used to allocate the N subbands with the specific bandwidth as in [3]. The temporal envelope from each band was extracted by half-wave rectification and low-pass filtering (fourth-order Butterworth) at either 50 or 500 Hz. The temporal envelope of each band was used to modulate a sinusoid generated at the center frequency of the analysis band. Amplitude-modulated sinusoids from all bands were summed and then normalized to yield the same root-mean-square amplitudes as the input signals. The training sets were used to train models. Maximum likelihood (ML) training was applied to estimate models. All synthesized testing sets were recognized with the system trained on signals with the same synthesis parameters.

2.1. Connected tone recognition task

The NUM-100A database consists of connected Mandarin Chinese digit strings [6]. The signals were recorded in a normal laboratory environment at an 8 kHz sampling rate and encoded with 16-bit linear PCM. The database included 8000 Mandarin digit strings produced by 50 males and 50 female speakers, and these digit strings included 1000 each of two-, three-, four-, six-, and seven-digit strings respectively, plus 2000 single digit utterances. Among the 8000 Mandarin digit strings, 7520 with different lengths were used for training, while the other 480 with different lengths were used for testing. A 25-ms Hamming window shifted with 10-ms steps and a pre-emphasis factor of 0.97 were used to obtain 39-dimensional feature vectors of 13 static coefficients (c0-c12) and first and second derivatives. Each of the 10 Mandarin digits was modeled by a whole hidden Markov model (HMM), each consisting of five states, eight Gaussian mixture components per state. In addition to the ten digit models, a silence model was constructed, consisting of three active states, eight Gaussian mixtures per state. Since testing data included only the 10 digits, the recognition process is to select one out of the 10 digits. The free-decoded error rates over all of the 500 clean testing utterances is 8.18%.

3. EXPERIMENTAL RESULTS

3.1. Effects of amplitude modulation bandwidth

Figure 1 shows a significant effect of the envelope filters in the connected tone recognition task. For the 500-Hz filtering condition, 1-, 2-, 3-, and 4-subband performance was greater compared to the 50-Hz filtering condition. These recognition

results contrast with previous results on a balanced database [5], where changing the bandwidth of the subband temporal envelope had no significant effect on accuracy [1].

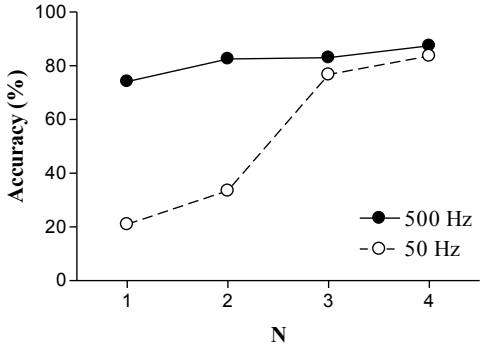

Figure 3: *Accuracy as a function of number of N subbands, with envelope filter frequency at 500 Hz* (•) *and at 50 Hz* (o).

3.2. Temporal-domain feature compensation

Feature compensation approaches reduce signal variability. This section evaluates adjusting the temporal distribution of testing and training signals to possess similar characteristics.

1) *Cepstral mean subtraction (CMS)*: normalizes the mean component of a feature sequence to remove DC components. CMS is performed by Eq. (1):

$$\hat{C}^i[t] = C^i[t] - \mu^i, t = 1, \dots, T, \quad (1)$$

where μ^i is the average of the sequence $C^i[t], t = 1, \dots, T$, and T is the number of frames.

2) *Cepstral mean and variance normalization (CMVN)*: in addition to normalizing the mean component as in Eq. (1), CMVN also normalizes the variance of the feature sequence:

$$\hat{C}^i[t] = \frac{C^i[t] - \mu^i}{\sigma^i}, t = 1, \dots, T, \quad (2)$$

where σ^i is the standard deviation of the sequence $C^i[t], t = 1, \dots, T$.

3.3. Temporal normalization of envelope

Table 1 shows feature compensation improved performance for conditions that evaluated the relative contributions of temporal envelope (2-50 Hz). For 50-Hz filtering conditions, CMS and CMVN reduced error rates when spectral information was limited. Cepstral normalization was also shown to reduce error rates in microphone array-based recognition [8] for SRS [1] with subband temporal envelope extraction filters using the same 50 Hz cut-off frequency. The Kullback-Leibler divergence between the global probability density functions revealed train/test mismatch reduction.

	1	2	3	4
Baseline	82.10	61.72	36.41	16.7
+CMS	80.14	*61.66	27.06	17.27
+CMVN	*79.04	66.55	*23.26	*16.35

Table 1: *Error rate (%) for N = [1, 2, 3, 4] with the envelope filter frequency at 50 Hz (o). Asterisks for best performance.*

3.4. Temporal normalization of periodicity

Table 2 shows feature compensation degraded performance for conditions that evaluated the relative contributions of periodicity (50-500 Hz). For 500-Hz filtering conditions, CMS or CMVN disrupted the complementary contributions of amplitude contour and periodicity [7].

	1	2	3	4
Baseline	*24.7	*14.68	*12.32	*10.09
+CMS	26.54	15.2	14.1	12.38
+CMVN	25.85	17.5	16.93	12.55

Table 2: *Error rate (%) for N = [1, 2, 3, 4] with the envelope filter frequency at 500 Hz (•). Asterisks for best performance.*

4. CONCLUSION

Periodicity is especially important for recognition of tones. Front-end processing should utilize alternative temporal filters to preserve tonal envelope. The construction of robust temporal features could benefit from deriving data-driven temporal filters using optimization criterion of principle component analysis (PCA) or minimum classification error [9]. Temporal feature statistics normalization [10] and subspace compression can be applied to wavelet streams with inverse discrete wavelet transform (DWT) reconstruction.

5. REFERENCES

[1] C.-T. Do, D. Pastor, and A. Goalic, "On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR," *IEEE Trans. On Audio, Speech, and Language Processing*, 18(5), pp. 1065-1068, 2010.

[2] C.-T. Do, D. Pastor, and A. Goalic, "A novel framework for noise robust ASR using cochlear implant-like spectrally reduced speech," *Speech Commun*., 51(1), pp. 119-133, 2012.

[3] P. Lin, F. Chen, S.-S. Wang, Y.-H. Lai, and Y. Tsao, "Automatic speech recognition with primarily temporal envelope information," *in Proc. INTERSPEECH,* pp. 476-480*,* 2014.

[4] R. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, 270, pp. 303-304, 1995.

[5] R. Leonard, "A database for speaker-independent digit recognition," *in Proc. IEEE ICASSP*, 9, pp. 328-331, 1984.

[6] [Online] Available: http://rocling.iis.sinica.edu.tw/

[7] Y-.Y, Kong, F.-G. Zeng, "Temporal and spectral cues in Mandarin tone recognition," *J. Acoust. Soc.*, 120(5), pp. 2830-2840, 2006.

[8] C.-T. Do, M.J. Taghizadeh, and P.N. Garner, "Combining cepstral normalization and cochlear implant-like speech processing for microphone array-based speech recognition," *in IEEE workshop on Spoken Language Technology (SLT)*, 2012.

[9] J.-W. Hung, and L.-S. Lee, "Optimization of temporal filters for constructing robust speech features in speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, 14(3), pp. 808-832, 2006.

[10] J.-W. Hung, H.-T. Fan, "Subband feature statistics normalization techniques based on discrete wavelet transform for robust speech recognition," *IEEE Signal Processing Letters*, 16(9), pp. 806-809, 2009.