A STUDY OF TONES AND TEMPO IN CONTINUOUS MANDARIN DIGIT STRINGS AND THEIR APPLICATION IN TELEPHONE QUALITY SPEECH RECOGNITION[1]

Chao Wang and Stephanie Seneff

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
{wangc,seneff}@sls.lcs.mit.edu

ABSTRACT

Prosodic cues (namely, fundamental frequency, energy and duration) provide important information for speech. For a tonal language such as Chinese, fundamental frequency (F_0) plays a critical role in characterizing tone as well, which is an essential phonemic feature. In this paper, we describe our work on duration and tone modeling for telephone-quality continuous Mandarin digits, and the application of these models to improve recognition. The duration modeling includes a speaking-rate normalization scheme. A novel F_0 extraction algorithm is developed, and parameters based on orthonormal decomposition of the F_0 contour are extracted for tone recognition. Context dependency is expressed by "tri-tone" models clustered into broad classes. A 20.0% error rate is achieved for four-tone classification. Over a baseline recognition performance of 5.1% word error rate, we achieve 31.4% error reduction with duration models, 23.5% error reduction with tone models, and 39.2% error reduction with duration and tone models combined.

1. INTRODUCTION

Prosody is mainly correlated with fundamental frequency (F_0), duration and energy features of a signal. It clearly conveys linguistic information about speech on various levels: sentence, word, syllable, etc. We are becoming increasingly interested in the use of prosodic aspects of speech to improve speech recognition. In this regard, we feel that continuous Mandarin digits form an excellent domain in which to carry out our initial study. First, digit strings have relatively simple sentence level prosodic structure, so we can focus on characterizing word level prosodic features as a first step. Second, F_0 plays a critical role in determining tones for Mandarin syllables, which is important to Mandarin recognition. Third, digits cover all four lexical tones in Mandarin; thus continuous digit strings provide an adequate domain in which to begin to study tones and their contextual effects. We plan to extend our study to more linguistically rich domains such as Mandarin GALAXY [4], in which spontaneous conversations were carried out between a subject and a computer for information seeking. It will be interesting to incorporate the phrase and sentence level features into prosodic modeling.

Mandarin digit recognition has been investigated by many researchers [5] [6]. However, we are unable to find reported results under similar conditions as ours (speaker-independent, continuous, and telephone-quality). Continuous digit recognition is more difficult for Mandarin than for English for several reasons:

1. Each digit is pronounced as a mono-syllable, and half of them have only sonorant sounds.
2. Three digits are pronounced as single vowels: "yi1[2]"(1), "er4"(2) and "wu3"(5). Segmentation of vowel vowel sequence such as "yi1 yi1" is difficult, especially due to frequent absence of glottal stops in continuous speech. The counter problem of vowel splitting also exists. This leads to high insertion and deletion error rates for these three digits.
3. Digits "yi1" and "wu3" also tend to be obscured in coarticulation with other digits, such as in "qi1(7) yi1", "liu4(6) wu3" and "jiu3(9) wu3", etc. This leads to even higher errors for "yi1" and "wu3".
4. There exist several confusing digit pairs, such as "er4 / ba1(8)", "liu4 / jiu3", and "liu4 / ling2(0)", etc., partially due to the poor quality of telephone speech.

In this paper, we develop a duration model and a tone model aimed at reducing these errors. Duration modeling is motivated by the observation that insertion and deletion errors generally produce unusually long or short hypothesized words, resulting in bad duration scores. In order for this simple strategy to be effective, it is advantageous to reduce the variances for the duration measurements, and we propose a normalization scheme to achieve this. Tone modeling is developed to better discriminate confusing candidates, which often differ in tone. We find that the tone model also helps reduce some insertion and deletion errors, i.e., the F_0 contour pattern is clearly different for "wu3" and "wu3 wu3", even though the spectral shape and duration in the two cases could be very similar. We applied both models to post-process the recognizer N-best list. They yielded substantial performance gains over a baseline system individually, with further improvement realized when the two were combined.

2. CORPUS

The corpus[3] was collected automatically by recording phone calls from native Chinese speakers, and the waveform was sampled at 8 kHz. A different list of 30 random phone numbers (containing 9 digits) and 30 random digit strings (containing 5-10 digits) was given to each participant, and the subjects were

[1]This research was supported by DARPA under contract N66001-96-C-8526, monitored through Naval Command, Control and Ocean Surveillance Center.

[2]Chinese pin-yin representation, augmented with tone.

[3]provided by Applied Language Technologies (ALTech) in Boston.

DATA SET	TRAIN	TEST
No. of strings	3923	355
No. of speakers	71	6

Table 1: Summary of the corpus.

instructed to read from the list in a naturally speaking way. Refer to Table 1 for a summary of the corpus. The average string length is 8.1 digits.

3. EXPERIMENTAL FRAMEWORK

The baseline recognizer is configured from the SUMMIT segment-based system [2]. There are 11 words in our vocabulary, including the alternative pronunciation "yao1" for digit "one". Chinese syllable initials and finals are chosen as the phone model units, with the addition of closure, inter-word pause, glottal stop, and nasal ending models, introduced by phonological rules. To fully exploit the small vocabulary size, digit-specific segment and boundary models are used. We achieved 5.1% word error rate (34.1% string error rate) on the test data with an A^* search.

We have found that the N-best list has great potential for improved performance. With a perfect post-selector, we could achieve less than 1% word error rate (7% string error rate) with a 10-best list. Thus, instead of incorporating the scores from duration and tone models into an early search stage, we decided to apply them in post-processing the 10-best outputs.

The post-processing scheme is similar to that proposed in [3]. For each word in an A^* path, the duration and/or tone scores are added to the total score, with the total adjustment normalized by the number of words (to avoid bias toward shorter strings). The N-best hypotheses are then resorted according to the adjusted total scores to give a new "best" sentence hypothesis. Context-dependent model scores can also be applied simply by converting a hypothesis into its context-dependent form, with context obtained from its surrounding hypotheses. In this way, an incorrect hypothesis is likely to result in bad scores both for its own wrong identity and as context for its neighbors. We also have a scaling factor to weight the scores contributed by duration and/or tone models, which can be optimized empirically based on recognition performance.

4. DURATION MODELING

Although logarithmic duration is already a feature in our *segment* model, it is likely to be poorly utilized in the principle component analysis, because of the high dimensionality of the segment feature vector, as well as the large variance due to, among other things, different speaking rate. Motivated by the observation that insertion and deletion errors generally produce unusually long or short hypothesized words, we isolate the *word* level duration as a separate feature to model. However, we only take the final part of a syllable as the "word" duration, with the belief that the syllable final is more stable and less subject to segmentation errors.

In order for the duration model to be more sensitive to insertion and deletion errors, it is critical to reduce the inherent duration variances caused by natural variations among different speakers. For simplicity, we assume that each digit in an utterance has roughly the same speaking rate. Therefore, we propose to normalize the duration of each word with respect to an estimated *sentence level* speaking rate. Given a phonetic alignment of an utterance, the speaking rate is estimated as

$$Speaking\ Rate = \frac{\mu_{DUR}(Words\ in\ a\ sentence)}{\mu_{DUR}(Words\ in\ corpus)}.$$

When computing the sentence mean, each word's duration is first normalized separately for that digit's mean duration to avoid bias caused by the content of the utterance (some digits tend to be intrinsically longer than others). However, the phonetic alignments for an utterance are not available in actual recognition. We developed a simple algorithm to estimate the speaking rate from the N-best list instead. We first obtain statistics of adjusted durations for all the finals in the N-best paths. After discarding anomalous tokens, the average of the remaining "reliable" candidates is then used to compute the speaking rate.

System Configuration	WER(%)	SER(%)
Baseline	5.1	34.1
+ Unorm. Dur.	3.9	26.8
+ Norm. Dur. w/ Est. Spk. Rate	3.5	24.8
+ Norm. Dur. w/ "True" Spk. Rate	2.9	20.3

Table 2: Comparison of recognition performance with different duration models. All conditions except the baseline have separate word-level duration models.

The duration distribution is modeled by mixtures of Gaussians. Duration scores from the classifier are passed on to post-process the N-best list, as described in the previous section. We obtained significant performance improvement over the baseline, as summarized in Table 2. However, the performance using speaking rate estimated from the N-best list is significantly worse than that using "true" speaking rate (estimated from forced alignments). We are continuing to refine the estimation algorithm to reduce this gap.

5. TONE MODELING

Tone recognition is an important part of Chinese speech recognition. It is a challenging research problem, due to (1) the difficulties of extracting fundamental frequency reliably, especially for telephone-quality speech, (2) the discontinuity of the parameter space due to the voiced/unvoiced dichotomy, as well as the need for normalization to account for the wide range of F_0 variations across speakers, and (3) the complex context dependencies of tone expression, mediated through tone sandhi rules.

5.1. F_0 Extraction

Fundamental frequency extraction is particularly difficult for telephone-quality speech, due especially to the fact that the fundamental is often weak or missing. To address this problem, we have developed a new pitch-extraction algorithm, which is based on the principle that harmonics will be spaced by the same amount on a *logarithmic* frequency scale regardless of the fundamental. More formally, if a signal has periodic peaks spaced at period P, then, on a logarithmic scale, the peaks will occur at $log(P)$, $log(P) + log(2)$, $log(P) + log(3)$, ..., etc. Thus the period P only affects the *position* of the first peak. We sample

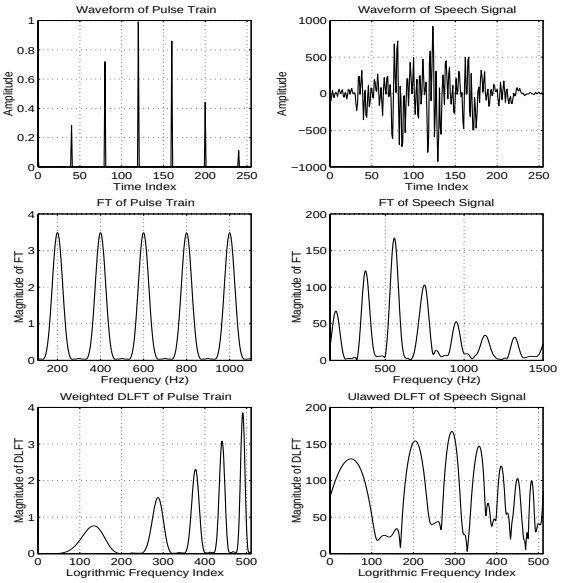

Figure 1: Windowed waveform, FT, and adjusted DLFT (refer to the text for details) for a pulse train and a speech signal.

a narrow band spectrum in the low-frequency region $[f_s, f_e]$ at linear intervals in the logarithmic frequency dimension. We define this representation as a *discrete logarithmic Fourier transform* (DLFT). Given a Hamming windowed speech signal $x(n)$ $(n = 0, 1, ..., N_w-1)$, the DLFT is computed as

$$X_i = \frac{1}{N_w} \sum_{n=0}^{N_w-1} x(n) e^{-j\omega_i n} \quad (i = 0, 1, \ldots, N-1),$$

where

$$\omega_i = 2\pi e^{(lnf_s + i \cdot dlnf)} \cdot T_s \quad (i = 0, 1, ..., N-1),$$

$$dlnf = (lnf_e - lnf_s)/(N-1),$$

N is the size of the DLFT, and T_s is the sampling period of the waveform. Figure 1 shows the waveform, Fourier transform and DLFT for a $200Hz$ pulse train and a voiced speech signal. The DLFT of the speech signal, sampled between 150 and 1500 Hz, has been normalized by μlaw conversion to flatten out the formant peaks. The *weighted* DLFT of the pulse train, which is used as a *template* for F_0 extraction as described later, was normalized such that each lobe has roughly equal area.

Because the spectral change due to vocal tract resonances should be relatively small between adjacent frames, we expect similar DLFT spectra except for an offset dependent on $log(F_0)$. A cross-correlation computation provides a *robust* estimate of the relative shift $\Delta log(F_0)$. We similarly compute a cross-correlation between each frame and the template, to obtain F_0 estimates. Both parameters are considered jointly in determining the F_0 value, taking into account continuity constraints.

Our current pitch extraction algorithm works as follows. Within a voiced region, provided either by a phonetic transcription or by a voiced/unvoiced decision algorithm, we first try to find the F_0 value for an anchor point, usually the first frame. All the frames in the region will each vote on the F_0 value for this anchor, based on its correlation with the template and the *cumulative* $\Delta log(F_0)$ from the anchor. The F_0 for the anchor is then taken as the median of all the votes, and F_0 values for the other frames are computed from the anchor F_0 and relative $\Delta log(F_0)$. We feel that this method can ensure a smooth F_0 contour within the voiced region because of the *robust* $\Delta log(F_0)$ estimation between adjacent frames. However, the F_0 value for the anchor frame could still have errors (when the majority of the frames make errors in their votes). We perform a sentence level smoothing to further correct potentially doubled/halved F_0 segments. It is possible that the F_0 contour is doubled/halved throughout the whole sentence. However, this type of error can be corrected by pitch normalization, and thus is less harmful to tone recognition.

We did not formally evaluate our pitch tracking algorithm. However, we feel that its performance in tone recognition is a valid indirect assessment criterion. We have manually examined the agreement between the extracted pitch contour and harmonic peaks in the DLFT spectrogram for many utterances, as shown in Figure 2, to ensure that the algorithm is performing well.

5.2. Tone Features

Tone is mainly dependent on the F_0 contour pattern, i.e., its average, slope, and curvatures. There are various ways to quantify these features in a segment-base system, by either fitting the F_0 contour with a certain type of function, or projecting it onto some basis functions. We have chosen the first four coefficients of the discrete Legendre transformation as our tone features, following the example of Chen & Wang [1], who describe the decomposition procedure in detail. They used the transformation for Mandarin pitch coding.

In a speaker-independent system, it is necessary to normalize the absolute F_0 with respect to the average over the entire utterance, to reduce across-speaker differences. We determined empirically whether to adjust by a ratio or a sum. Our data indicate that the ratio gives smaller variances for the pitch-related features we have chosen; thus the Legendre coefficients are *scaled* according to the average F_0 of the utterance.

Normalized duration, as introduced in the previous section, is also included as a basic segment feature. The duration feature does not contribute significantly to tone discrimination, but it is essential to limit insertion and deletion errors in recognition.

A principle component analysis is applied to our five-dimensional tone feature vector, and mixtures of diagonal Gaussians are used to model the distributions.

5.3. Context

Each of the four Mandarin lexical tones has a basic F_0 pattern [7]. However, the actual pitch contour for a tone can vary dramatically in different tonal contexts. In order to characterize contextual effects systematically, we performed a clustering experiment on "tri-tone" models. Besides the four tones, we also included a "blank" context to represent sentence start, sentence end and long inter-word pauses, resulting in 100 "tri-tone" models ($5 \times 4 \times 5$).

The cluster tree shows that the models divide naturally into four

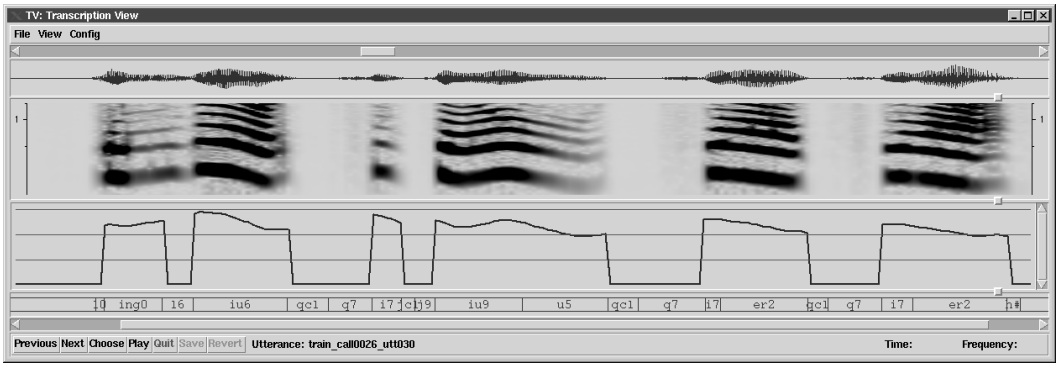

Figure 2: Waveform, $\mu law ed$ DLFT spectrogram, extracted F_0 contour, and phonetic transcription for the Mandarin digit string "ling2 liu4 qi1 jiu3 wu3 qi1 er4 qi1 er4"(067957272).

No. of Classes	4	21	51	100
Error Rate (%)	20.0	20.2	20.0	20.7

Table 3: Classification error rate for tone models with different number of "tri-tone" classes.

System Configuration	**WER(%)**	**SER(%)**
Baseline	5.1	34.1
+ Duration Model	3.5	24.8
+ "Tri-tone" Model (21 classes)	3.9	25.9
+ Both Models	3.1	21.1

Table 4: Summary of recognition results for different systems.

major categories, each corresponding to a lexical tone class. An exception is the well-known tone-sandhi rule that third tone becomes second tone when preceding third tone. We can obtain merged tone classes of different levels of detail from the cluster tree, by varying the distance threshold. We experimented with 4, 21, 51 and 100 context-dependent tone classes. Refer to Table 3 for a summary of four-tone classification results. It seems that these tone models have roughly the same performance, except for a slightly higher error rate for the 100-class model, probably due to under-training. Notice that the 4-class context-dependent tone model is different from a context-independent 4-tone model in that the tone-sandhi rule for third tone is taken into consideration. The context-independent model has 22.7% error rate.

6. WORD RECOGNITION PERFORMANCE

Table 4 summarizes the recognition performance with the duration model, the tone model, and both models. Experiments show that the tone model with 21 classes has a small gain over the others in recognition. This might be because it is more sensitive to contexts than the 4-class model and more robust than the 51-class model in post-processing the N-best list. It is encouraging to see that, when used in conjunction, the duration and tone models yielded further performance gains.

7. SUMMARY

We have built a duration model and a tone model for the continuous Mandarin digit task, and obtained substantial performance gains by applying them to post-process the N-best outputs of a baseline system. This demonstrates the potential of using prosodic features for improved speech recognition. We also developed a new robust pitch extraction algorithm, particularly suitable for telephone quality speech. In this study, we have assumed the prosody to be stable throughout the utterance. In fact, even digit strings could have various prosodic structures, such as pauses in phone numbers and long digit strings. It will be interesting to incorporate these factors into prosodic modeling. Eventually, we plan to extend this work to more linguistically rich domains such as Mandarin GALAXY, where the influence of sentence level prosody may play an important role.

8. REFERENCES

1. S. Chen and Y. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Transactions on Communications*, Vol. 38, No. 9, pp. 1317-1320, September 1990.

2. J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-fased speech recognition," *Proc. ICSLP '96*, Philadelphia, PA, USA, pp. 2277-2280.

3. B. Serridge, *Context-Dependent Modeling in a Segment-Based Speech Recognition System*, S.M. thesis, MIT, 1997.

4. C. Wang, J. Glass, H. Meng, J. Polifroni, S. Seneff, and V. Zue, "YINHE: A Mandarin Chinese version of the GALAXY system," *Proc. Eurospeech '97*, Rhodes, Greece, pp. 351-354.

5. C. Wang, S. Tang, D. Liang, H. Chen, and C. Tang, "Methods for combining the information of various features in speech recognition," in *Acta Acustica*, vol.22, no.2, pp. 111-115, 1997.

6. J. Wang, C. Wu, C. Huang, and J. Lee, "Integrating neural nets and one-stage dynamic programming for speaker independent continuous Mandarin digit recognition," *Proc. ICASSP '91*, Toronto, Ont, Canada, Vol. 1, pp. 69-72.

7. Y. Wang, J. Shieh, and S. Chen, "Tone recognition of continuous Mandarin speech based on Hidden Markov Model," *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 8, No. 1, pp. 233-246, 1994.