ISCA Archive
http://www.isca-speech.org/archive

7th International Conference on Spoken
Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

IMPROVING PARAMETRIC TRAJECTORY MODELING BY INTEGRATION OF PITCH AND TONE INFORMATION *

Yiyan Zhang, Wenju Liu, Bo Xu and Huayun Zhang

National Laboratory of Pattern Recognition, Chinese Academy of Sciences,
P.O.Box 2728, Beijing, P. R. China, 100080
yyzhang,lwj,xubo,hyzhang@nlpr.ia.ac.cn

ABSTRACT

This paper presents the application of pitch/tone information to improve Parametric Trajectory Modeling (PTM). To simulate the trajectory of pitch in a segment in PTM recognizer, we in fact get its corresponding tone information. From another point of view, tone is a segmental feature and PTM has the excellent framework to incorporate it. So we here introduce the "soft" and "hard" integration methods to combine them with the base modeling. In the experiment of Mandarin digit classification, they get 22.87% and 33.54% error reduction respectively, and integration of both them together obtains 38.72% error reduction.

1. INTRODUCTION

It has been shown in the empirical observations that input cepstra of speech signals are not stationary in general but vary with time across the duration of the speech segment. This explicitly illuminates the dynamics of features and the existence of trajectories in speech signal. Better performance for speech recognition would be achieved by taking into account these information. Now a variety of models broadly classified as segment models have been addressed to explore these useful information [1]. Through time-warping (or re-sampling), segment models can be encoded by a fixed-length representation for the variable-length observed segment. It can be thought of as an underlying trajectory and thus features' dynamic characteristic can be denoted. So far ever used cepstra in segment models include short-time energy and MFCC. Excellent potential information of pitch and its contours----tone, however hasn't been considered sufficiently yet.

Compared with other features (e.g. short-time energy, MFCC), pitch is embedded with dynamic quality. And to simulate its trajectory in a segment, we get the corresponding tone. As is known, Mandarin is a monosyllable-structured language used by the most speakers in the world. Different from other languages, it is with unique lexical tones. Statistically almost 30% words are respective homonyms when ignoring tone [5], but the ratio is tremendously reduced if tone information is considered. So information of tone plays a very important role in disambiguating the syllables. Therefore pitch/tone is very useful information in segment modes for Mandarin recognition.

The remainder of the paper is organized as follows: In section 2, we address a segment model----parametric trajectory modeling as the recognizing framework. Next, in section 3, we describe methods to integrate pitch/tone information with baseline modeling and introduce pitch extraction algorithm. And finally, experiment results and conclusions are given in section 4.

2. PARAMETRIC TRAJECTORY MODELING

With many possible distribution assumptions that represent feature dynamics, several segment alternatives can be outlined, among which includes the parametric trajectory modeling (PTM).

Gish and Ng introduced the parametric trajectory modeling with the normalized time processing [2]. The model represents the dynamics of features in a segment by means of parameterization through constant, linear, or higher order polynomial trajectory. A speech segment Y with N frames and D dimensions can be modeled as:

$$C = ZB + E \qquad (1)$$

Where Z is an $N \times (R+1)$ design matrix to normalize different length of segments uniformly between 0 and 1. B is a $(R+1) \times D$ trajectory parameter matrix and E is a residual error matrix. We reduce training cost by taking advantage of an assumption that the covariance is diagonal in the segment. R is the polynomial order.

3. TONE INFORMATION

Pitch is the frequency of vocal cords vibration when speaking vowels. It changes with time in either a single syllable or continuous speech and its variable trajectory is called tone. For the main classed tone1, tone2, tone3 and tone4 each is with its trajectory figure that is similar to their signal Mandarin representation "¯", "´", "ˇ" and "`".

The idea of the parametric trajectory modeling is to simulate the feature variation in a segment. Compared with MFCC and energy, pitch deserves its own dynamic characteristics, and above all its trajectory is tone which is an independent and important component in phonetics. So the perfect applications with pitch and tone will be demonstrated in the following.

3.1. Pitch/tone's alignment within base modeling

How to integrate pitch/tone information into base modeling is a significant problem to be handled. According to the model and feature's characteristics, two methods in this paper have been

* This work is supported by Grants No. 60172055 from Chinese National Science Foundation and No. 4002010 from Beijing Government Science Foundation.

proposed----the soft integration and the hard integration [3][4][5].

The parametric trajectory modeling models each speech segment as a curve (or collection of curves) in cepstral space. The type of curve (namely trajectory) is specified by R, which is thus far considered no more than 2: $R=0$ for constant, $y=b_{0j}$; $R=1$ for linear, $y=b_{0,j}+b_{1,j}t$; $R=2$ for quadratic, $y=b_{0,j}+b_{1,j}t+b_{2,j}t^2$, where j ($j=1,\cdots,D$) represents some dimension, and t is normalized time, as illustrated in Fig. 1.

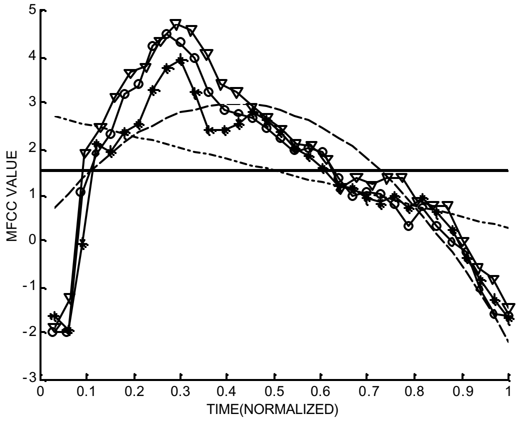

Fig. 1: Cepstral feature C_1 for three training data of "dong1", each represented by "$\triangleleft$" "$*$" or "o"; and the Fit of Constant (solid), Linear (dash-dotted), and Quadratic (dashed) trajectories for the features.

Through above analysis, we conclude that: The essence of parametric trajectory modeling is, using R degree polynomial function to simulate the trajectory of features for each dimension in the parametric space. If we take pitch as the 14th dimension of the feature, just as short-time energy and MFCC do, its trajectory is in fact the tone curve of the syllable, which is significant to speech recognition.

Above is one method to integrate pitch information into acoustic modeling. We call it in this paper the soft integration.

For all segment models, there is a key aspect which we consider to be useful----it can incorporate the information acquired from segmental feature measurements [1]. As to the segmental features, we mean features measured over the time span of a segment. In the work presented formerly, the most useful and broadly investigated segmental feature is segmental duration.

By its definition, tone is a fine feature across time span. We can take it as a segmental-level feature. So another method is generated to integrate tone information into acoustic modeling, which we call it the hard integration. Now the decoding formula for the recognition process is modified as

$$\hat{\alpha}=\arg\max_{\alpha}\{\log P(Y\mid\alpha)+\lambda\bullet P(T\mid\alpha)\}, \qquad (2)$$

Where $P(T\mid\alpha)$ is the tone score of a speech segment with tone T given model α , and λ is a weighting factor to match dynamics ranges of traditional $\log P(Y\mid\alpha)$ and $P(T\mid\alpha)$. It is difficult to determine it. In section 4.2.2 we introduce a method to avoid selection of this value.

Assuming that with the help of some advanced knowledge, we've segmented the speech very accurately to the syllables. The remaining problem is focused on how to decide tone attribute and then align it within the base modeling score. These will be introduced at length in the experiment part.

3.2. Pitch extraction

To use pitch/tone information efficiently, pitch extraction should be required to be more precise. Therefore an improved algorithm [6] is given in this paper as the following:

Step 1: Calculating the auto-correlation coefficients.

Ideal auto-correlation function is with the following form:

$$R(m)=E\{x(n)x(n+m)\}=\lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}x(n)x(n+m)\,. \qquad (3)$$

Because of the constriction of short time analysis, speech signal is truncated to the frames of 10~20ms:

$$x_i(n)=x(n)*w(n-iN)$$

among which $w(n)$ is the window function.

So the short-time auto-correlation function is usually denoted as:

$$\begin{aligned}\hat{R}(m)&=\frac{1}{N}\sum_{n=0}^{N-1-|m|}x_i(n)x_i(n+m)\\&=\frac{1}{N}\sum_{n=0}^{N-1-|m|}x(n)w(n-iN)x(n+m)w(n+m-iN)\end{aligned} \qquad (4)$$

There is apparently warp between $\hat{R}(m)$ and $R(m)$.

$$\begin{aligned}&E\{\hat{R}(m)\}\\&=\frac{1}{N}E\left\{\sum_{n=0}^{N-1-|m|}x(n)w(n-iN)x(n+m)w(n+m-iN)\right\}\\&=\frac{1}{N}\sum_{n=0}^{N-1-|m|}E\{x(n)x(n+m)\}w(n-iN)w(n+m-iN) \qquad (5)\\&=\frac{R(m)}{N}\sum_{n=0}^{N-1-|m|}w(n-iN)w(n+m-iN)\\&=R(m)R_w(m)\end{aligned}$$

where $R_w(n)$ is the auto-correlation function of $w(n)$. We can get the desired $R(m)$ from $\hat{R}(m)$:

$$R(m)=\frac{E\{\hat{R}(m)\}}{R_w(m)} \qquad (6)$$

then a new improved method is derived to estimate $R(m)$:

$$R^*(m)=\frac{\hat{R}(m)/\hat{R}(0)}{R_w(m)/R_w(0)}\quad , \qquad (7)$$

where $R_w(m)$ is the auto-correlation function of $w(n)$. $R^*(m)$ is called the normalized zero-distortion auto-correlation estimation. As another approximation of $R(m)$, however it is more precise than $\hat{R}(m)$.

Step 2: Select top-5 frequencies corresponding to top-5 auto-correlation coefficients as the pitch candidates.

Step 3: Use Dynamic Programming (DP) algorithm to balance the minimization of frequencies variation and maximization of auto-correlation coefficients.

We use about 30s telephone data to measure performance of the algorithm. Assuming the accuracy of manual extraction is 100%, the algorithm result is shown as follows [6].

Testing data / Error type	male	female
Constant/vowel decision	3.77%	4.04%
Frequency warp	0.87%	1.01%

Table 1: The accuracy of pitch extraction

4. EXPERIMENT RESULTS

4.1. Data corpus description

A continuous digits database collected by our National Laboratory of Pattern Recognition in 1995 is used for training and testing. Our data corpus includes 55 males, and each people has 80 utterances. The length of each utterance varies from 1 to 7. In our experiments, we take 40 males' data for training and the remaining 15's data for testing.

In order to make our data corpus contain as many phenomena as possible, we designed our data corpus deliberately. The digit string in our data corpus has the following statistical properties:

1) All digits are designed with roughly the same probability to be uttered (Each person has utterance 31 “0”, 27 “1”, 29 “2”, 30 “3”, 30 “4”, 32 “5”, 30 “6”, 28 “7”, 29 “8”, 29 “9”, 26 “Y”, where “ Y” represents another commonly pronunciation of “ 1” in Mandarin);

2) All digits connection are considered and balanced;

3) Every digit's positions in the string are balanced;

For each digit, we build up a word model independently. There are 11 Mono-word models totally.

We use viterbi algorithm to segment speech data with the help of text transcription and develop the model case with these pre-segmented input data. All the following experiments are carried out on the same training set and testing set.

4.2. Results and analysis

Two methods have been introduced to integrate pitch/tone information into the baseline modeling Following is the particular processing.

4.2.1. The soft integration

In this case, pitch is integrated into the feature vector as the 14th component.

Here we first compare PTM with HMM. The HMM platform is Palm-PC command recognition system [7]. Its model unit is isolated word (in our experiment is digital syllable), with 8 states and 16 mixtures. The PTM platform uses the same model unit, 13MFCC, 16 mixtures, which is referred as the baseline. We can see from Table 2 that our PTM has much more precise modeling ability. And it even achieves higher accuracy with 13MFCC because it can fully express the feature's dynamics.

Table 2 is the soft pitch integration result. Utt. Wise Norm means pitch normalization based on the whole utterance. Without smoothing the result degrades slightly. This is because only vowel has pitch and the consonant part is set to zero. If we calculate parameters and simulate trajectory under this condition, several frames of zero will lead to severe distortion. So we use a simple smoothing strategy----replacing zero by one. The accuracy improves greatly after smoothing.

HMM (13MFCC+△+△△)	96.26%
HMM (13MFCC)	88.41%
Baseline (PTM,13MFCC)	96.72%
Pitch +Utt. Wise Norm	96.70%
Pitch +Utt. Wise Norm + smoothing	97.47%

Table 2: Accuracy improvement using pitch information

4.2.2. The hard integration

In this case, tone information is taken as the segmental-level feature. This method follows two steps.

Step 1. Tone attribute judgement

In a segment, we may judge tone attribute by its shape.

While curve emulating, we focus our work on the pre-processing, specially treating the beginning and ending sect of a syllable. This is due to the following reasons. Firstly, severe co-articulation in continuous speech exits and tone may be affected seriously by its left and right circumstances. Secondly, in the typical tone figures (tone1, tone2, tone3, tone4), there remains a characteristic property [8]. That is, the curves universally ascending at begin (approximately 60ms) and descending at the end (approximately 40~50ms), which are respectively referred as head plying and tail falling. It is indeed harmful when distinguishing tone. According to above facts, we discard different frames' pitch value at both edges. Experiments show that the left circumstances in continuous speech and heading plying in itself have much more ambiguous effect on tone. The best performance is obtained when discarding 2 frames at begin and 1 frame in the end.

We then use Linear Least Square Method to simulate tone and record the slope and absolute sum of the errors as parameters. With different measurement rules, the input data's tone attribute can be described. We can make one judgement for three cases: 1) it is explicitly one of the main classed tone1, tone2, tone3 and tone4. 2) with the linear method, the slope of tone1 and tone3 is nearly in the same scope. And sometimes the absolute sum of the errors is not small enough or big enough for us to make a explicit judgement just as in the above case. So we refer it as fuzzy tone, which may be tone1 or tone3. 3) the parameters do not satisfy the rigorous term and we make no judgement.

Tone property	meaning	flag
Classed tone	ˉ	1
	ˊ	2
	ˇ	3
	ˋ	4
Fuzzy tone	ˉ or ˇ	13
no tone	No judgement	-1

Table 3: Tone attributes

The result of tone attribute judgement is shown in Table 4. Study shows that the error is mainly because of the affects of the co-articulation in continuous speech and head plying and tail falling, which makes some digits' tone trajectory aberrant.

Clarity degree	Correct ratio	Wrong ratio
Classed tone	49.85%	14.27%
Fuzzy tone	10.89%	4.36%
No tone	20.63%	

Table 4: Accuracy of tone attributes decision

Step 2. Align two separate stream scores

After getting the tone attribute, we should calculate its score and then combine it with the baseline modeling. Here we didn't compute tone score separately, but considered it as a fraction of the acoustic likelihood. What we need to do is how to match their range, namely, to decide the coefficient of fraction.

The tone flag of the recognizing digit is denoted as the signal I and that of the model digit by J. According to the relationship between I and J, we define different **"Visibility",** as shown in table 5.

Satisfying conditions	Visibility degree
I=J	2
I=13 && J=1||J=3	1
I= -1	0

Table 5: Tone's visibility

Different visibility should be given different coefficients. The one belonging to degree 0 is certainly zero, and the other two need to decide, among which the one belonging to degree 2 is with most importance. Here we define three models:

visibility	2	1
Model1	1/5	1/15
Model2	1/13	1/39
Model3	1/14	1/42

Table 6: Coefficients for different tone attributes

In model1, we choose coefficients based upon experience; in model2 and model3, coefficient is set to reciprocal of the dimension number (13 and 14 respectively); All the coefficients for degree 1 are approximately third of the corresponding value for visibility 2. Table 7 shows the results for each model.

baseline	96.72%
Model1	97.18%
Model2	97.82%
Model3	97.82%

Table 7: Accuracy for different score match

4.2.3. The total integration

Here we use both pitch as 14th feature and tone as segmental-level feature at the same time. We get best performance 97.99% with model3.

baseline		96.72%
14th pitch feature +Tone score	Model1	97.49%
	Model2	97.90%
	Model3	97.99%

Table 8: Integrated performance

In Mandarin connected digit recognition, some digit pairs are easily confusable, for example, "2" and "8", "6" and "9", "6" and "Y". This is partially because they are similar in commonly used MFCC. Since we introduce pitch and tone information here, the confusions can be alleviated to some extent. Table 9 accounts for the conclusion in detail.

	"2" and "8"	"6" and "9"	"6" and "Y"	total errors / ratio
baseline	48	16	21	158 / 53.80%
Pitch +tone + model3	18	9	7	97 / 35.05%

Table 9: Number of digit pairs' confusion

We see that the confusion is degraded smartly.

So we conclude that the separated application of two methods and their integration all improve the performance greatly, and pitch/tone are very useful features in parametric trajectory modeling.

5. CONCLUSIONS

In this paper we present pitch/tone as typical dynamical features. To fully explore their role in the segment models, two methods----the soft and hard integration are used for combing them with acoustic modeling. From all these experiments, we point out: 1) Pitch as 14th feature and tone as segmental-level feature are very helpful for segment models. 2) Pitch normalization, tone attribute judgement and matching coefficients need further investigation to make consistent improvement of the performance.

6. REFERENCES

[1] M. Ostendorf, V.V.Digilakis and O.A.Kimball, "From HMM's to segment models: a unifies view of stochastic modeling for speech recognition," in *IEEE Trans on Speech and Audio Processing*, vol.4, no.5, 1996.

[2] Gish, H. & Ng, K. "A segmental speech model with application to word spotting," *Proc. of ICASSP'93*, Minneapolis, U.S.A., vol.Ⅱ, pp. 447-450, 1993.

[3] Jason Davenport, Richard Schwartz, Long Nguyen, "Towards a robust real-time decoder", *Proc. of ICASSP'99*, pp2409-2412.

[4] Bo Xu, Sheng Gao, Hua Wu, Taiyi Huang, "Integration tone information in continuous Mandarin recognition", *Proc. of ISSPIS,99*, Guangzhou, P.R.China.

[5] Eric Chang, et al, "Large vocabulary Mandarin speech recognition with different approaches in modeling tones", *Proc. of ICSLP'00*, vol 2, pp983-986.

[6] HuaYun Zhang, ZhaoBing Han and Bo Xu, "Speech recognition in telephone speech translation", *Proc. of NCMMSC6*, pp235-238, 2001, P.R.China, (in Chinese).

[7] YongGang Deng, Bo Xu and TaiYi Huang, "Palm-PC speech recognition algorithm and its realization", *Computer Application and Development*, vol. 37, No. 8, 2000, (in Chinese).

[8] XingJun Yang, HuiSheng Chi, *Digital Processing for Speech Signal*, China Electronic Industrial Publication, July 1998, (in Chinese).