INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 7, 115–128, 2004
© 2004 Kluwer Academic Publishers. Manufactured in The Netherlands.

Tone Modeling for Continuous Mandarin Speech Recognition

YANG CAO
Nokia Research Center, China
yang.1.cao@nokia.com

SHUWU ZHANG, TAIYI HUANG AND BO XU
National Laboratory of Pattern Recognition, Chinese Academy of Sciences
swzhang@hitic.ia.ac.cn
huang@nlpr.ia.ac.cn
xubo@hitic.ia.ac.cn

Abstract. Tone study is very important for Mandarin speech recognition. In this paper, a Mixture Stochastic Polynomial Tone Model (MSPTM) is proposed for tone modeling in continuous Mandarin speech. In this model the pitch contour, main representative of tone pattern, is described as a mixed stochastic trajectory. The mean trajectory is represented by a polynomial function of normalized time while the variance is time varying. Effective training and tone recognition algorithms were developed. The experimental results based on the proposed MSPTM showed 40.7% tone recognition error rate reduction relative to the traditional Hidden Markov Model (HMM) tone model. We also present a decision tree based approach to learning the tone pattern variation in continuous speech. The phonetic and linguistic factors that may affect the tone patterns were taken into consideration while constructing the tree. After the tree was established, 28 different tone patterns were obtained. We found that in addition to the tone of the neighboring syllable, Consonant/Vowel type of the syllable and the position of the syllable in the utterance also made important contributions to tone pattern variations in continuous speech. Finally, a new approach of integrating tone information into the search process at word level is discussed. Experiments on continuous Mandarin speech recognition showed that the new tone model and tone information integration method were efficient, achieving a 16.2% relative character error rate reduction.

Keywords: tone modeling, continuous speech recognition, Mandarin speech recognition

1. Introduction

One of the most important characteristics of the Chinese language is its tonal nature in which the same syllable with different tones represents different characters. In fact, there are at least 6763 commonly used Chinese characters (Chinese GB codes) and only 417 phonologically valid syllables in Mandarin Chinese without tone. How do those 417 syllables represent 6763 different characters? As a tonal language, Mandarin Chinese basically has four lexical tones and one neutral tone. A total number of 1302 different syllables can be composed by combining those tones and syllables. Then, those 6763 characters can share the pronunciations of 1302 syllables in what are called homophones. From this viewpoint, tone is a crucial factor for Mandarin speech recognition.

What is the nature of tone? The acoustic nature of tone lies in the variation of fundamental frequency (F_0). The four lexical tones in isolated syllables can be characterized mainly in terms of the shapes of their F_0 contours as: high level, rising, falling-rising and falling. Therefore, F_0 contour is the most crucial characteristic of tones. Research has shown that the duration of tones is also very important (Lin, 1998).

It is clear that tone classification in isolated syllables is less difficult because tone patterns are simple and stable in this condition. However, in the case of continuous speech, tone pattern classification is very difficult, because under this condition tone patterns are subject to variations. For example, the pronunciation of Tone 5 is highly dependent on its context so that its pattern is relatively arbitrary.

Therefore, there are two basic problems associated with tone classification in continuous Mandarin speech. The first one is how to build a reasonable mathematical model for tone patterns. The second one is how to characterize tone variations in continuous speech.

For the first problem, traditionally, there are two categories of tone modeling approaches commonly used for continuous Mandarin today. One is based on HMM, which is the mainstream of tone modeling. HMM-based methods have the advantage of fully-developed training algorithms and the convenience of integration with acoustic HMM models (Wang et al., 1997; Huang and Seide, 2000). However, the performance of this type of tone modeling method is not remarkable (Zhao et al., 1997; Wang et al., 1994).

Another commonly used tone model is based on neural networks (NN) (Chang et al., 1972; Chen and Wang, 1995; Chen et al., 1998). Different network structures like Multiple Layer Preceptor, Recurrent Network, and so on were utilized in NN-based models. NN-based systems work like a black box, in which the context dependency is taken into consideration and solved implicitly. On the one hand, you don't need to worry about context dependency at all; on the other hand, the result can't explain explicitly how context affected tone pattern variation and cannot be used to guide acoustic decoding procedures.

In additional to HMM and NN, the feature-based method also has been applied for tone modeling (Wang and Seneff, 1998). It gives a good tone recognition performance about 80% in digit-string recognition.

In this paper, we describe a new tone modeling approach based on stochastic polynomials, named Mixture Stochastic Polynomial Tone Model (MSPTM), to model tone features. In studying the F_0 contours in continuous Mandarin speech, we found that the third order polynomial was precise enough to characterize the F_0 contour. Considering the stochastic property of speech signals and the variations among different speakers, we utilized a statistical method to model the tone patterns (Cao et al., 2000). In the Stochastic Polynomial Tone Model (SPTM), the F_0 contour is represented as a stochastic trajectory. The mean trajectory is parameterized by a polynomial curve while the variance is time varying. Efficient training and recognition algorithms were developed. Compared with existing HMM or Neural Network tone models, the SPTM model gives a directly geometrical presentation of tone patterns and it can handle the problem of pitch value missing naturally, thus making the model more robust. In order to enhance the model characterization, we further extended the SPTM to the mixture polynomial tone model (MSPTM). The experimental results showed that the tone recognition error rate decreased by 40.7% compared with the traditional HMM tone model.

Regarding the second problem, most current analyses of tone pattern variations derive from qualitative observations. In order to overcome those shortcomings, we decided to investigate tone pattern variation in continuous Mandarin speech through the stochastic clustering method. The decision tree was chosen as our clustering method because it is a data-driven method that can incorporate expert knowledge. While constructing the decision tree, besides neighboring tones, many other factors were considered, including syllable position in the word and Consonant/Vowel type of the syllable, which were not utilized in conventional recognition systems. The decision tree was set up on a large corpus. After the tree was constructed, 28 different tone patterns were acquired. The results revealed that many phonetic and linguistic factors other than tone of neighboring syllable, affected tone variation patterns in continuous Mandarin speech. Considering all those factors mentioned above, a new context-dependent tone model was built.

How to utilize tone information in continuous Mandarin speech recognition is another important problem. Several methods have been proposed in the past few years. Some methods produced very good results (e.g., Huang and Seide, 2000). In Wang et al. (1997) and Huang and Seide (2000), tone recognition was separated from syllable recognition, while in Chen et al. (1997) and Wong and Chang (2001), tone was treated as an attribute of the phoneme, and pitch was integrated into the feature vectors. The former method cannot guide *N*-best candidates or word lattice generation, so it cannot recover any acoustic search error, and the separated decoding procedure may cause mismatches between tones and syllables. The latter method appears to be a natural tonal problem solution, but it produces too many acoustic models when considering the tone coarticulation effect. In this paper, an approach to incorporate a tone contribution score at the word level is proposed. By using the proposed context-dependent tone method generated by the decision tree, the present method integrates tone information into *N*-best or word lattices generation directly. Experiments on continuous Mandarin speech recognition have shown that the new tone model is very helpful, achieving a 16.2% relative recognition error rate reduction.

Section 2 presents the proposed SPTM method. In Section 3, an improvement of SPTM is proposed (Mixture SPTM). In Section 4, we discuss the method to learn tone pattern variations in continuous speech via decision tree clustering. The incorporation of tone information into the search process is described in Section 5. Some experimental results are reported in Section 6, and the final conclusion is presented in Section 7.

2. Stochastic Polynomial Tone Model (SPTM)

The Stochastic Polynomial Tone model consists of two parts: the pitch contour model and the duration model.

2.1. *Definition of Pitch Contour Model*

For tone pattern α, we assume that its pitch (F_0) sequence is $F = \{f_0, f_1 \dots f_L\}$where f_i represents the F_0 value at time frame i. In SPTM, this sequence is represented by an R-th order trajectory model as follows:

$$f_i = \mu(i) + \varepsilon(i) = \sum_{k=0}^{R} b_k^\alpha\,(i/L)^k + \varepsilon(i), \qquad (2.1)$$

here, $\mu(i)$ represents the mean at the time frame i, ε is white noise with zero mean and variance $v(i)$, and $B = \{b_k^\alpha\}, k = 0, 1 \dots R$ are the polynomial coefficients. The key to this model is how to specify the variance $v(i)$. Because the variance of the F_0 contour varies with the location in a syllable, in our approach a monotonic non-decreasing time warping mapping T_L is applied to split the syllable into M segments, while the variances of F_0 are fixed within each segment and vary across segments. Formally, $v(i) = V^\alpha_{T_L(l)}$ and

$$T_L : i \to m, m \in \{1 \dots M\}. \qquad (2.2)$$

The likelihood of a F_0 sequence F, given that it is generated by a tone pattern α, can be expressed as:

$$\begin{aligned} p_f(F \mid \alpha) &= \prod_{t=0}^{L} N(f_t - \mu^\alpha(t), v^\alpha(t)) \\ &= \prod_{t=0}^{L} N\left(f_t - \sum_{k=0}^{R} b_k^\alpha (t/L)^k, V^\alpha_{T_L(t)}\right), \end{aligned} \qquad (2.3)$$

here, $N(\ )$ means the normal distribution function.

2.2. *Definition of Duration Model*

Various types of duration modeling have been discussed by many researchers (e.g. Russell and Moore, 1985; Lee et al., 1998). Most of them are parameterized methods. Gaussian, Poisson and γ distribution functions all have been tried, but none of them can model adequately the complex property of duration. In order to solve this problem, the context-dependent duration model was proposed (Lee et al., 1998), but that model increases the system complexity significantly. Finite mixtures are flexible and powerful probabilistic modeling tools (Jain et al., 2000), and Gaussian mixture is the most popular one. Gaussian mixture is able to approximate arbitrary probability density functions (pdf's) (Hastie and Tibshirani, 1996), this makes it well suited for modeling complex class-conditional pdf's. We, thus, used it for the duration model. In this model, the likelihood of length L, given that it is generated by tone pattern α, can be expressed as:

$$p_l(L \mid \alpha) = \sum_{i=1}^{C} C_i^\alpha N\left(L - \mu_i^\alpha, v_i^\alpha\right), \qquad (2.4)$$

here, C_i^α is the weight coefficient of mixture component i, μ_i^α is the mean of the mixture component i, and $v_i^\alpha$is the variance of the mixture component i.

Based on the previous definition, the joint probability of a sample $F = \{f_0, f_1 \dots f_L\}$ and a model α is:

$$p(F \mid \alpha) = p_f(F \mid \alpha) \times p_l(L \mid \alpha)^\eta, \qquad (2.5)$$

here, the weight η is introduced to match the different dynamic ranges of p_f and p_l.

2.3. *Recognition Algorithm*

Suppose there are a total of k tone models C_i, $i = 1 \dots k$, and each model is characterized by parameter setsΘ^i. If an observing pitch sequence $F = \{f_0, f_1 \dots f_L\}$ is given, according to Bayesian decision rules, the classifying result can be written as:

$$C(F) = C_i \text{ if } p(C_i \mid F) = \max_j p(C_j \mid F). \tag{2.6}$$

Then we can define:

$$g(F \mid C_i) = \log p(C_i \mid F). \tag{2.7}$$

By omitting the constant for all patterns, g can be rewritten as:

$$\begin{aligned} g(F \mid C_i) &= 2\log P(C_i) + 2\eta \log(p_l(L \mid C_i)) \\ &\quad - \sum_{t=0}^{L} \left[\log\left(V^i_{T_L(t)}\right) + \left(f_t - \sum_{k=0}^{R} b^i_k (t/L)^k \right) \Big/ V^i_{T_L(t)} \right]. \end{aligned} \tag{2.8}$$

The decision rule now becomes:

$$C(F) = C_i \text{ if } g(F \mid C_i) = \max_j g(F \mid C_j). \tag{2.9}$$

2.4. *Training Algorithm*

Let $\Psi = \{X_1 \dots X_N\} X_i = \{x_{i,0} \dots x_{i,L_i}\}$ be a set of training data that belongs to the model α. The parameters of the pitch contour model and the duration model are estimated separately.

2.4.1. Duration Model. The training data are split into two parts. One part is used for training, and the other part serves as the cross-validation data. The standard Expectation Maximization (EM) (Dempster et al., 1977) algorithm is used to estimate duration model parameters. The mixture number k is determined via maximum likelihood of the cross-validation data. In our system, the search space of k is: $1 \le k \le 8$.

2.4.2. Pitch Contour Model. The parameters of the pitch contour model can be estimated via Maximum Likelihood rules. In order to further optimize the model parameters, the minimum classification error (MCE) training was also developed.

(A) Maximum Likelihood Rules. Given the training data $\Psi = \{X_1 \dots X_N\}$, the model parameters $\theta^\alpha = \{b^\alpha, V^\alpha\}$ can be estimated via the Maximum Likelihood rules. That is, we wish to find model parameter θ^*:

$$\theta^* = \arg\max_\theta p_f(\Psi \mid \theta). \tag{2.10}$$

The above formula can be solved by the following equations:

$$\begin{cases} \dfrac{\partial p_f(\Psi \mid \theta)}{\partial b_k} = 0 & k = 0 \dots R \\ \dfrac{\partial p_f(\Psi \mid \theta)}{\partial V_m} = 0 & m = 1 \dots M \end{cases} \tag{2.11}$$

Expanding Eqs. (2.11) yields

$$\begin{aligned} &\sum_{k=0}^{R} b_k \sum_{i=1}^{N} \sum_{t=0}^{L_i} [t/L_i]^{r+k} \Big/ V_{T_{L_i}(t)} \\ &\quad = \sum_{i=1}^{N} \sum_{t=0}^{L_i} \frac{x_{i,t}}{V_{T_{L_i}(t)}} [t/L_i]^r, \end{aligned} \tag{2.12}$$

for all $r = 0 \dots R$.

$$\begin{aligned} V_m = &\left(\sum_{i=1}^{N} \sum_{t=0}^{L_i} \delta(m, T_{L_i}(t)) \right. \\ &\left. \times \left[x_{i,t} - \sum_{k=0}^{R} b_i [t/L_i]^k \right]^2 \right) \Big/ \sum_{i=1}^{N} \sum_{t=0}^{L_i} \delta(m, T_{L_i}(t)), \end{aligned} \tag{2.13}$$

for all $m = 1 \dots M$.

Here, δ represents Kronecker delta. By solving the system of Eqs. (2.12) and (2.13), we obtain the required parameters $B^\alpha$and V^α. However, Eqs. (2.12) and (2.13) are nonlinear equations, so it is very difficult to find an analytical expression. Therefore, we developed the following iteration algorithm to estimate the parameters:

Algorithm 1:

1. Let $V^0_m = V^0_1$, for $m = 2 \dots M$.
2. Replacing V_m in Eq. (2.12) with $V^{(i-1)}_m$, yields $b^{(i)}_r$;
3. Replacing b_r in equation (2.13) with $b^{(i)}_r$, yields $V^{(i)}_m$;
4. Calculate the probability of training data with respect to $b^{(i)}_r$, $V^{(i)}_m$ and denote it as $P^{(i)}$;
5. If $\log(P^{(i)}) - \log(P^{(i-1)}) < \varepsilon$, go to step 7;
6. $i = i + 1$, go to step 2;
7. $b^\alpha_r = b^{(i)}_r, V^\alpha_m = V^{(i)}_m, r = 0 \dots R, m = 1 \dots M$, Termination.

(B) MCE Training. MCE is an effective discriminative training algorithm (Juang and Katagiri, 1992; Juang et al., 1997). It is able to correctly discriminate the observations for best classification results rather than to fit the model parameters to the data (ML).

Assume there are K distinct tone models; formula (2.8) can be used as the discriminant function for model i. However, the first two parts can be ignored in MCE training, so the function can be simplified as:

$$
\begin{aligned}
g_i(X,\theta) = -\sum_{j=1}^{M}\sum_{t=0}^{L}\delta(j, T_L(t))\Bigg[&\log\left(V_j^i\right) \\
&+ \left[x_t - \sum_{k=0}^{R}(t/L)^k\right]^2 \Bigg/ V_j^i\Bigg]. \qquad (2.14)
\end{aligned}
$$

The *class misclassification measure* can be expressed as:

$$
\begin{aligned}
d_i(X) &= -g_i(X,\theta) \\
&\quad + \log\left(\frac{1}{k-1}\sum_{j,j\neq i}\exp(g_j(x,\theta)\eta)\right)^{1/\eta}. \qquad (2.15)
\end{aligned}
$$

Then, the classification error function $L(\theta)$ can be obtained. By using the Generalized Probabilistic Descent algorithm, we can obtain the final iteration formula.

In order to maintain the constraint of the model ($V_{i,m} > 0$), the following parametric transform is used during parameter adaptation:

$$
V_m^i \to \tilde{V}_m^i = \log\left(V_m^i\right). \qquad (2.16)
$$

For the training sample $X_m \in C_i$in the training set, the discriminative adjustment is:

$$
\begin{aligned}
b_k^i(n+1) = b_k^i(n) &- 2\varepsilon_n\lambda l_i(d_i(X_m,n)) \\
&\times (1 - l_i(d_i(X_m,n))) \\
&* \sum_{j=1}^{m}\sum_{t=0}^{L_m}\delta\big(j, T_{L_m}(t)\big)(t/L_m)^k \\
&\times \left(\sum_{r=0}^{R} b_r^i(n)(t/L_m)^r - x_{m,t}\right) \Bigg/ V_j^i(n) \\
\tilde{V}_k^i(n+1)& \\
= \tilde{V}_k^i(n) &- \varepsilon_n\lambda l_i(d_i(X_m,n))(1 - l_i(d_i(X_m,n))) \\
&* \sum_{t=0}^{L_m}\delta\big(k, T_{L_m}(t)\big)\Bigg[1 - \left[x_{m,t} - \sum_{r=0}^{R} b_r^i(t/L_m)^r\right]^2 \\
&\times \exp\big(-\tilde{V}_k^i(n)\big)\Bigg] \qquad (2.17)
\end{aligned}
$$

Here,

$$
l(d) = \frac{1}{1+\exp(-\lambda d)}.
$$

For parameters of competing models, a similar process can be derived to adjust the parameters.

2.5. *Case of Pitch Value Missing*

In continuous Mandarin speech, there are some weak voiced parts that often appear to be unvoiced, such as the center part of Tone 3. Some unexpected zero points often exist in the estimation sequences of pitch. This problem may be more serious in practical speech recognition systems. Moreover, as the recognition results cannot be completely error free, some unvoiced parts may be taken as voiced while segmenting syllables. In traditional HMM or NN tone models, due to the constraint of the models, this problem is usually tackled by extracting F_0 with a lower threshold followed by a smoothing process, which might unfortunately result in incorrect F_0 contours. However, the proposed stochastic tone model is tolerant of pitch value missing.

Assuming $f_i = 0$ in the sequence F, because it provides no information, we can define $p(f_i \mid \alpha) = 1$ for any tone pattern α. The formula can be changed to:

$$
\begin{aligned}
g(F \mid C_i) = 2\log P(C_i) &+ 2\eta\log(p_l(L \mid C_i)) \\
&- \sum_{t=0}^{L} u(f_i)\Bigg[\log\left(V_{T_L(t)}^i\right) \\
&+ \left[f_i - \sum_{k=0}^{R} b_k^i(t/L)^k\right] \Bigg/ V_{T_L(t)}^i\Bigg], \qquad (2.18)
\end{aligned}
$$

here

$$
u(x) = \begin{cases} 1x > 0 \\ 0x \le 0 \end{cases}.
$$

The training algorithm can be modified similarly. So, in our model, only the reliable F_0 is utilized in the training and recognition process, and the effect of pitch value missing is eliminated intrinsically, which ensures the robust-ness and precision of the model.

3. Mixture Stochastic Polynomial Tone Model (MSPTM)

When checking the model parameters of SPTM, we find that the variance value is still not small, which means mismatches exist between training data and models. This suggests that one stochastic trajectory cannot represent the complex distributions of tone patterns. As we know, different speakers have different pronunciation characteristics; the same tone pronounced by different speakers has different distributions even in the same syllable that has exactly the same context. One way to solve this problem is to represent the tone pattern with several polynomial curves. Finite mixtures probabilistic modeling was applied to this problem: a Mixture Stochastic Polynomial Tone model was proposed. There are two different mixture strategies: mixture at the output distribution level and mixture at the trajectory level. Mixture at the trajectory level can be viewed as the mixture of several stochastic curves. It has more trajectory constraints than mixture at the density level, which seems more suitable for describing tone patterns in continuous speech. Thus, MSPTM mixture at the trajectory level is used.

3.1. Model Definition

As stated above, in the mixture polynomial tone model, our goal is to model the pitch contour of a tone pattern using a mixture of stochastic polynomial curves. More specifically, for a pitch sequence $F = \{f_0, f_1 \ldots f_L\}$, we model the probability of F given tone pattern α as:

$$P_f(F \mid \alpha) = \sum_{j=1}^{Q} p(c_j \mid \alpha) p_f(F \mid c_j, \alpha), \qquad (3.1)$$

here, for each of the Q mixture components c_j, $p_f(F \mid c_j, \alpha)$ gives the probability of the whole pitch sequence conditioned on that component, and $p(c_j \mid \alpha)$ is the mixture weight. As we mentioned above, each component is represented by a stochastic polynomial model, assuming its parameters are:

$\theta_j = \{b_{j,k}, V_{j,m}\}\{k = 0 \ldots R, m = 1 \ldots M\}$, according to the definition of SPTM in Section 2.1. (The superscript α of parameters has been omitted for brevity.) Then, the probability of tone sequence F derived from tone pattern α is:

$$P_f(F \mid \theta) = \sum_{j=1}^{Q} \lambda_j p_f(F \mid \theta_j) = \sum_{j=1}^{Q} \lambda_j \prod_{t=0}^{L} N \times \left(f_t - \sum_{k=0}^{R} b_{j,k}(t/L)^k, V_{j,T_L(t)} \right), \qquad (3.2)$$

here, λ_j is the mixture weight.

The same Gaussian mixture duration model is included in MSPTM. Thus, in MSPTM, the joint probability of a sample F and the model α can be computed as:

$$p(F \mid \alpha) = P_f(F \mid \alpha) \times p_l(L \mid \alpha)^{\eta}. \qquad (3.3)$$

3.2. Training Algorithm

The clustering algorithm and the EM algorithm were applied to estimate the MSPTM parameters.

3.2.1. Clustering Method. In the clustering algorithm, the training data $\Psi = \{X_n\}(n = 1 \ldots N)$ for tone pattern α are partitioned into Q clusters using the k-means algorithm. The k-means algorithm requires a definition of the distance measure between a datum X and cluster m. In our approach, the *log* likelihood of a sample X and cluster m is used as the distance measure. Assuming the parameters of cluster i are $\{b_i, V_i\}$, the distance measure can be expressed as:

$$D(X, i) = \sum_{t=0}^{L} \left[(x - \mu_{i,t})^2 / V_{i,T_L(t)} + \log V_{i,T_L(t)} \right]. \qquad (3.4)$$

Estimates of $\{b_i, V_i\}$ can be obtained in the same way as in section (2.4), using the data assigned to cluster i, and λ_i is calculated as the relative frequency of the data as:

$$\lambda_i = N_i / N, \qquad (3.5)$$

here, N_i is the number of data assigned to cluster i.

3.2.2. EM Method. The Expectation Maximization (EM) algorithm is an efficient tool by which to estimate parameters of mixture models. In the EM algorithm, we assume the training data $\Psi = \{X_1 \ldots X_N\}$ are *incomplete* data, with the missing part $Y = \{y_1, y_2, \ldots y_N\}$

indicates which of the Q components generated X_i, if it was the mth component, then $y_i = m$. Thus the log-likelihood of the complete data with model parameters θ is:

$$L(\Psi, Y \mid \theta) = \sum_{i=1}^{N} \lambda_{y_i} p_{y_i}(X_i \mid \theta). \tag{3.6}$$

The Q-Function is:

$$Q(\theta, \theta^g) = E(L(\Psi, Y \mid \theta) \mid \Psi, \theta^g), \tag{3.7}$$

here, θ^g means current estimation of θ.

Maximization of the Q-Function will obtain the updated parameters estimate θ^{g+1}.

$$\theta^{g+1} = \arg\max_{\theta} Q(\theta, \theta^g). \tag{3.8}$$

Solving the function (3.8), we obtain:

$$\lambda_l(g+1) = \frac{1}{N} \sum_{i=1}^{N} p_f(l \mid X_i, \theta^g), \tag{3.9}$$

here,

$$p_f(l \mid X_i, \theta^g) = \frac{\lambda_l^g \prod_{t=0}^{L_i} p_f\left(x_{i,t} \mid \theta_l^g\right)}{P_f(X_i \mid \theta^g)}, \tag{3.10}$$

and

$$\begin{aligned} &\sum_{k=0}^{R} \sum_{i=1}^{N} p_f(i \mid X_i, \theta^g) \sum_{t=0}^{L_i} (t/L_i)^{k+j} \big/ V_{l,T_{L_i}(t)} \\ &\quad = \sum_{i=1}^{N} p_f(i \mid X_i, \theta^g) \sum_{t=0}^{L_i} x_{i,t} (t/L_i)^j \big/ V_{l,T_{L_i}(t)}, \end{aligned} \tag{3.11}$$

for all $j = 0 \ldots Rl = 1 \ldots Q$.

$$\begin{aligned} &V_{l,m} \\ &= \left[\sum_{i=1}^{N} p_f(l \mid X_i, \theta^g) \sum_{t=0}^{L_i} \delta(m, T_{L_i}(t))(x_{i,t} - \mu_{i,t})^2 \right] \Bigg/ \\ &\left[\sum_{i=1}^{N} p_f(l \mid X_i, \theta^g) \sum_{t=0}^{L_i} \delta(m, T_{L_i}(t)) \right], \end{aligned} \tag{3.12}$$

for all $m = 1 \ldots M \; l = 1 \ldots Q$.

Comparing Eqs. (3.11) and (3.12) with Eqs. (2.12) and (2.13), it is obvious that Eqs. (3.11) and (3.12) are the weighted forms of (2.12) and (2.13), respectively, with the weights $p_f(l \mid F_i, \theta^g)$. Therefore, we can use algorithm 1 in Section 2.4 to estimate $\{b\}$ and $\{V\}$. The procedure is:

1. Calculate $p_f(l \mid X_i, \theta^g) l = 1..Q$, $\lambda_l(g+1) \quad l = 1..Q$ based on $b_{l,j}(g)$ and $V_{l,m}(g)$,
2. Use algorithm 1 to estimate $b_{l,j}(g+1)$ and $V_{l,m}(g+1)$ for all mixtures l,
3. Determine if the stopping rules have been met; if not, go to step 1.

As we know, the EM algorithm is critically dependent on initialization. So, in our system, we use the cluster method to obtain the initial model, then, the EM algorithm is applied to optimize the model parameters.

4. Tone Pattern Clustering

Before constructing the tree, we must consider the following aspects: input data, the question set, the evaluation function and stopping criteria. By optimizing these parameters, we can obtain a satisfactory tone classification in continuous speech. The binary tree was used in our approach.

4.1. Input Data

Since the tone of a syllable is characterized mainly by its F_0 contour, and duration is also related to it, the features we were used were normalized pitch sequences and duration.

4.2. The Question Set

We included many factors that may affect the variance of tone patterns in the question set. It consists mainly of the following items:

1. Tone of neighboring syllables. Based on previous research, we know that tone of adjacent neighboring syllables was the main factor for tone pattern variations in continuous Mandarin speech. Our approach ignored the influence of farther neighboring syllables and focused on the effect of adjacent syllables. We designed five questions for the tone

Table 1. Six types of Consonant.

1	{m, n, l, r, “null”}
2	{f, s, sh, x, h}
3	{b, d, g}
4	{p, t, k}
5	{j, zh, z}
6	{c, ch, q}

Table 2. Fourteen types of Vowels.

7	{a, ia, ua}
8	{o, uo}
9	{e, ie, ve}
10	{i}
11	{v}
12	{er}
13	{ai, uai}
14	{ei, uei}
15	{ao, iao}
16	{ou, iou}
17	{an, ian, uan, van}
18	{ang, iang, uang}
19	{eng, ong, ing, iong}
20	{en , in, uen, vn}

of the left syllable and the same number of questions for the tone of the right syllable. Some researchers proposed that that the influence of Tone 1 and Tone 2 (or, Tone 3 and Tone 4) on the following tones is very similar; as do Tone 1 and Tone 4 (also, Tone 2 and Tone 3) on the preceding tone. We include those four tone-combination questions in our question sets. There are a total of 14 questions about the tone of neighboring syllables.

2. Syllable position in the utterance. We have observed that when a syllable is a phrase boundary, its tone association suffers different co-articulation effects from those syllables inside the phrase. Therefore, the syllable position problem must be included in the question sets. We designed some questions about syllable position, such as if the current syllable is a monosyllabic word, if the current syllable is at the beginning of the word, etc. There are a total of six questions about syllable position.
3. Consonant/Vowel type of the syllable. Consonant/vowel type is another important factor for tone pattern variations. For example, voiced consonants have pitch value, while unvoiced consonants have no pitch value. The Consonant/Vowel types considered in our approach are listed in Tables 1 and 2.

Based on the above three factors, we defined a total of 40 questions for decision tree based tone pattern clustering.

4.3. Evaluation Function

The evaluation function is a measure of the purity of the data; it is used to determine the validity of node splitting. In the speech recognition area, the entropy function and the function of the probabilistic sum are commonly chosen as evaluation functions for node splitting. As discussed in Section 2, we employ the Stochastic Polynomial Tone Model (SPTM) to characterize tone patterns. Therefore, the product of the probabilistic values of all the samples in a node can be used as a measure of how well the model fits the data at that node; if the data within the node are similar to each other, that value will be large. As discussed in Section 2.1, for F_0 sequence $F = \{f_0 \ldots f_L\}$, its probability derived from tone model $\Theta = \{B, V, T\}$ is

$$P(F \mid \Theta) = p_f(F \mid B, V)p_l(L \mid T)^{\eta}. \tag{4.1}$$

Assuming that the samples belonging to node α are $\{F_i\}(i = 1 \ldots k)$ and the model corresponding to that node is $\Theta(\alpha)$, then, the probability is:

$$P(\alpha) = \prod_{i=1}^{k} P(F_i \mid \Theta(\alpha)).$$

We define $L(\alpha) = -2 * \log(P(\alpha))$ as the evaluation function. Now, if a question q splits the data at node n into two sub-nodes n_l and n_r, the outcome of this split is:

$$m(n, q) = L(n) - L(n_l) - L(n_r). \tag{4.2}$$

So, $m(n, q)$ is a measure of the improvement in purity as a result of this split by question q.

The procedure of evaluating $L(\alpha)$ when creating the decision tree is:

1. Estimate SPTM parameters from the samples $\{F^i\}(i = 1 \ldots K)$ belonging to node α; and
2. Compute evaluation function $L(\alpha)$.

4.4. Stopping Criteria

Stopping criteria determine when to stop the node splitting. In our system, we used a very simple stopping criterion: if the outcome of the best question splitting at a node n is less than the threshold, or the number of the samples at a node falls below the threshold, we stop the split of this node and designate it as a leaf node. The thresholds are selected empirically.

5. Integrating Tone Information into Continuous Speech Recognition

As previously mentioned, the presented tone models are dependent on syllable position in the word and Consonant/Vowel type of the syllable, so tone decoding cannot be separated from other decoding processes. One possible way to use tone information in speech recognition is to rescore N-best candidates, but it is not an efficient way, because the tone information cannot be incorporated directly into N-best generation under this strategy. For utilizing the tone information in speech recognition efficiently, we have developed an algorithm to integrate tone cue into the decoding process directly. The basic idea is to add the tone contribution score for those words that reach the last state and begin to expand to the next word. For example, in our system, we add tone score in the procedure of word lattice generation, that is, when a word in one path reaches its end state and begins to determine if it can be added to the word lattice. The process can be illustrated as follows:

At every time frame t, for each word W_i that reaches its end state and begins to extend to the successor tree, we can obtain its score as:

$$\begin{aligned} P(W_i, t) = {} & P(W_p, t_p) + K_b B(W_i) \\ & + \mathit{TScore}(Wi, t, t_p) + A(W_i, t, t_p), \end{aligned} \quad (5.1)$$

here, W_p represents W_i's best previous word in the current path, $t_p$is W_p's end time, $P(W_i, t_k)$ represents the total score for word W_i which ends at t_k in the current path, $B(W)$ stands for the language model score associated with word W in current path, $A(W_i, t, t_p)$ represents word W_i's total acoustic score from time frame t_p to t, and $\mathit{TScore}(W_i, t, t_p)$ represents the tone model score for word W_i which starts at time frame t_p and ends at time frame t.

The tone score is calculated as follows:

$$\mathit{TScore}(W_i, t, t_p) = k_t(t - t_p)/L(t, t_p)T(W_i, t, t_p), \quad (5.2)$$

here, k_t represents the tone information weight that was empirically selected, and $L(t, t_p)$ is the total voiced phone length between time t and t_p. Because tone information exists only in the voiced part of a syllable, we do this normalization to make all the score at the same time frame, have the same normalized length, so they can be compared to each other. $T(W, t, t_p)$ is the cumulative tone score of the current word, which can be calculated as follows:

$$T(W, t, t_p) = \sum_{n=1}^{N_w} V(tone_n, t_s, t_E), \quad (5.3)$$

here, $V(n, t_s, t_E)$ denotes the probability score of model n assigned to the data starting at time frame t_s and ending at time frame t_E. N_w denotes the total number of syllables with tone associations in word W. Since we search on the lexical tree, we know each word's exact corresponding Consonant/Vowel series and each syllable's tone association. We also know acoustic environment of each tone, so we can map all syllables' tones to the corresponding tone model except the last one because we do not know its right neighbor's tone. To manage this problem, we can try all tones of the first syllable in words that can possibly follow the current word, and select the largest likelihood score as the current result. The only remaining parameters of the formula (5.2) and (5.3) are the starting time and ending time of each voiced phone, which can be acquired easily by recording some additional information during the search process. After this calculation, we can integrate tone information into the decoding process naturally.

Another issue that deserved consideration was computing complexity. Because there are fewer dimensional features, tone decoding is much faster than phoneme decoding. With respect to integrating tone information at the word level, a more precise match score is achieved, so a strong pruning strategy can be chosen to make the search more efficient.

6. Experiments

6.1. Experimental Conditions

Experiments were conducted based on the Chinese continuous speech database of the national “863” project, which was designed to cover as many coarticulation effects as possible. We chose a subset consisting of eight males and eight females, yielding a total of 9,740 sentences and 110,696 syllables as training data. In total there are 23,048 syllables with Tone 1, 27,128 syllables with Tone 2, 16,372 syllables with Tone 3, 37,712 syllables with Tone 4 and 6,472 syllables with Tone 5. In the training procedure, each sentence was divided into Consonant/Vowel segments by viterbi alignment. An auto-correlation pitch determination process with a dynamic programming post-processing algorithm was applied for pitch detection. Then, the pitch contour of each sentential speech sample was normalized by averaged pitch value of that sentence to reduce interspeaker variability. Data from other four male and four female speakers was chosen as the testing set. There were a total of 4,370 sentences and 55,348 syllables in the testing database.

In the following experiments, the order of the stochastic polynomial model was set to three. The time-warping transform T_L, given the variance region number M was set as:

$$T_L(x) = \begin{cases} 1 & x < 2 \\ \cdots & \\ i & (i-2)\dfrac{L-3}{N-2} < (x-1) \le (i-1)\dfrac{L-3}{N-2} \\ \cdots & \\ N & x > L-2 \end{cases} . \tag{6.1}$$

As we know, at the beginning and end of a syllable in continuous speech, there are transitions. At transitions, pitch contour is unstable. In T_L, the first and last lines refer to transitions. The remaining steady segment was split evenly.

A five-state, three output continuous density HMM was used for the HMM tone model, and each output was represented with 4-mixture Guassians. The features of HMM are normalized pitch and energy and their first-order and second-order derivatives.

6.2. Statistical Learning Result

In our experiments, we built one decision tree for each tone pattern. Each leaf node of the tree defines a special tone variation pattern. At the same time, the questions of the leaf node describe the context condition of this tone pattern. By analyzing the questions of a leaf node, we can find what influences the tone variations the most and how it works. In the clustering procedure, the order of SPTM is set to three; and the variance number is four. When the number of samples in a node is less than 300 or the increase in the evaluation function is less than the threshold, the node will stop splitting. We obtained 28 tone variation patterns from decision tree clustering. The main results are listed in Table 3.

By analyzing the questions in the leaf nodes of the decision tree, we found that questions related to the tone

Table 3. Main results of decision tree clustering.

Tone 1	1	Syllable at the head of word
	2	Syllable not at the head; Sonant
	3	Syllable not at the head; not sonant
Tone 2	4	Left = silence
	5	Left ≠ silence; Right = Tone 3
	6	Left ≠ silence; Right = Tone 2
	7	Left = Tone 3; Right ≠ Tone 2 or 3
	8	Left ≠ silence or Tone 3; Right ≠ Tone 1, 2 or 3; Consonant
	9	Left ≠ silence or Tone 3; Right = Tone 1; Consonant
	10	Left ≠ silence or Tone 3; Right ≠ Tone 2, 3 or 4; Sonant
	11	Left ≠ silence or Tone 3; Right = Tone 4; Sonant
	12	Right = Tone 3
	13	Left = Tone 4; Right ≠ Tone 3; Consonant
	14	Left = Tone 4; Right ≠ Tone 3; Sonant
	15	Left = silence; Right ≠ Tone 3; Consonant
Tone 3	16	Left ≠ silence or Tone 4; Right ≠ Tone 3; Sonant
	17	Left = silence; Right ≠ Tone 3; Sonant
	18	Left ≠ silence or Tone-4; Right ≠ Tone 3; Consonant {i, ia, ua}
	19	Left ≠ silence or Tone-4, Right ≠ Tone 3; Consonant except {i, ia, ua}
	20	Left = Tone 3
	21	Left = Tone 4; Right ≠ Tone 4
Tone 4	22	Left = Tone 4; Right = Tone 4
	23	Left ≠ Tone 3, 4 or silence
	24	Left = silence
	25	Left = Tone 1
	26	Left = Tone 2
Tone 5	27	Left = Tone 3
	28	Left = Tone 4

of neighboring syllables constructed 62%; questions related to Consonant/Vowel types of the syllable and syllable position took up to 19%, respectively.

From the above result, it is clear that tone context information is the major factor that affects the distribution of tone variations in continuous speech. We can conclude that the tone context of the neighboring syllable together with the Consonant/Vowel type of the syllable and the syllable position in the utterance made important contributions to tone pattern variations in continuous speech. In the following experiments, all tone models are context-dependent models based on the result of the decision tree.

6.3. SPTM Recognition Result

6.3.1. Comparison with HMM Tone Model. We firstly evaluated the tone recognition performance of the proposed Stochastic Polynomial Tone Model and the context-dependent HMM Tone Model. Because the HMM-based approach does not take into account the duration information, we ignored Duration Model p_l in this comparison, and the variance region number was set to four. In all of the following experiments, the average recognition rate is the ratio of the total number of correctly recognized tones to the total number of recognition occurrences. Because the number of occurrences for different tones is not equal, it is not the mean of the recognition rate of each tone.

The experiment results (Table 4) showed that the tone recognition performance of the SPTM model is significantly higher than that of the HMM-based approach. The average relative error rate reduction is 8.29%. The improvements in Tone 1, Tone 2 and Tone 3 are obvious. Note that the training algorithm of HMM has been well-developed, whereas the parameters of SPTM have not yet been optimized. We further tested the influence of factors such as number of variance number, the Duration Model, and training algorithms in the following experiments.

Table 4. Performance comparison of SPTM and HMM.

	Tone recognition rate (%)					
Models	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
HMM	73.41	64.18	59.13	68.77	54.14	66.33
SPTM	76.64	69.46	66.05	68.39	52.90	69.12
Error reduction	12.15	17.29	16.93	−1.22	−2.70	8.29

6.3.2. Experiment on Number of Variance. Based on the above comparison, we tested the model's performance with different numbers of variance or, in other words, when splitting the syllables into different numbers of segments.

Table 5 shows that the change of the numbers of variance had little influence on the performance. We also found that for each tone pattern, the variance of the steady parts varied very little with the different numbers of variance. This fact indicated that the proposed time warping function T_L fit the real distribution well. Based on the analysis above, we therefore fixed the variance number at four in the following experiments.

6.3.3. Experiment on Duration Model. The duration of a tone has proved to be helpful in tone perception. We thus evaluated the duration model for different weighting factors, as shown in Table 6.

From Table 7, we can see that incorporating duration information improves tone recognition performance and that duration information differentially influences different tones.

6.3.4. Experiment on MCE Training. To improve the system performance, we employed the Minimum

Table 5. Performance of different variance number.

Variance number	Tone recognition rate (%)					
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
3	76.64	69.41	66.07	68.18	52.90	69.04
4	76.64	69.46	66.05	68.39	52.90	69.12
5	76.69	69.38	66.16	68.25	52.87	69.08

Table 6. Result of the duration model.

Duration weight	Tone recognition rate (%)					
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
0	76.64	69.46	66.05	68.39	52.90	69.12
0.5	77.52	69.56	66.42	68.82	53.03	69.53
1.0	78.42	70.27	66.10	69.18	53.09	69.98
2.0*	78.59	70.38	66.00	69.71	53.32	70.33
3.0	78.34	70.21	66.15	69.25	52.26	69.93

Table 7. Result of MCE training.

	Tone recognition rate (%)					
Models	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
Baseline	76.64	69.46	66.05	68.39	52.90	69.12
Baseline + DM	78.59	70.38	66.00	69.71	53.30	70.33
MCE + DM	79.67	73.78	70.15	74.95	60.26	74.08
Error reduction	12.97	14.15	12.08	20.07	15.63	16.06

*DM means duration model.

Classification Error (MCE) algorithm to optimize model parameters. In the training procedure, the number of competing classes was set to four and the whole process was iterated five times.

In Table 7, the baseline system refers to the Stochastic Polynomial Tone Model. Experimental results showed that MCE training could generate more optimal models and achieved a 16.06% reduction in recognition error rate.

Based on the above experiments, we can finally make a comparison between the HMM tone model and the optimized SPTM model.

As shown in Table 8, optimized SPTM method achieved a recognition error rate reduction of 23% compared with the HMM tone model. We further take into account the number of parameters of those two methods into account. As mentioned above, the feature size of the HMM tone model is six, the number of parameters for a four mixture Gaussian output is 51. One HMM model with three outputs has 153 parameters, at least. The SPTM has only 32 parameters, at most. The number of parameters in the SPTM is less than one fourth that of the HMM tone model, which means it requires less training data to achieve a robust parameters estimation. Furthermore, the advantage of recognition speed is obvious. The time cost for recognizing tones of eight speakers by the SPTM is 25 seconds, and that of the HMM is 374 seconds.

Overall, the SPTM has the advantage in computational complexity, complexity of model parameters and recognition correction over the HMM tone model. It is obvious that the SPTM tone model is more suitable than the HMM tone model for describing tone patterns in continuous speech.

Table 8. Comparison of HMM and SPTM tone models.

	Recognition result (%)					
Model	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
HMM	73.41	64.18	59.13	68.77	54.14	66.33
SPTM	79.67	73.78	70.15	74.95	60.26	74.08
Error reduction	23.54	26.80	26.96	45.73	19.79	23.02

Table 9. Influence of mixture number.

Mixture	Tone recognition rate (%)					
number	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
2	84.31	81.58	69.37	76.29	62.30	77.42
3	86.00	82.90	70.02	76.60	63.54	78.37
4	86.35	83.65	72.15	77.25	65.27	79.26
5*	83.36	82.81	70.09	83.05	63.88	80.02
6	82.60	81.71	68.10	81.59	60.91	78.63
8	81.74	80.88	66.28	83.09	50.90	77.91

6.4. MSPTM Recognition Result

In our approach, K-Means clustering is used to initialize the parameters of the MSPTM, and the EM algorithm is used to optimize them.

We conducted some experiments to assess the influence of mixture numbers. The result is shown in Table 9.

The recognition accuracy is increased significantly when the MSPTM is applied. When the mixture number is set to five, it achieves the best performance. Compared to the SPTM, the error rate reduction is 22.91%.

The influence of different speakers on the HMM, SPTM and MSPTM was investigated. The result is shown in Table 10.

The 'Distortion' of the above table is the variance of recognition rates. From the result, the distortion of the MSPTM proved to be the lowest of the three. It showed that the proposed MSPTM was less sensitive to speaker diversity.

Finally, the overall performance comparison of the HMM, SPTM, and MSPTM is shown in Table 11. From the result, we can see that the MSPTM achieved the best performance. Compared to the HMM model, the error rate reduction is more than 40%.

6.5. Integrating Tone Information into Continuous Speech Recognition

The effectiveness of our framework for integrating tone information into the search process then was examined.

Table 10. The comparison on the alleviation of speaker diversity.

(%)	HMM	SPTM	MSPTM	Error reduction
Speaker 1	68.66	78.21	81.37	40.56
Speaker 2	69.56	76.53	80.83	37.02
Speaker 3	67.50	76.94	79.57	37.14
Speaker 4	66.75	74.61	81.55	44.51
Speaker 5	68.76	74.55	80.73	38.32
Speaker 6	66.10	73.00	79.90	40.71
Speaker 7	63.50	71.34	77.69	38.88
Speaker 8	60.46	68.12	78.51	45.56
Average	66.33	74.08	80.02	40.66
Distortion	2.87	3.08	1.29	

Table 11. Comparison of different tone models.

	Tone recognition rate (%)					
Model	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Average
HMM	73.41	64.18	59.13	68.77	53.14	66.33
SPTM	79.67	73.78	70.15	74.95	60.26	74.08
MSPTM	83.36	82.82	70.09	83.05	63.90	80.02
Error reduction	36.59	52.04	26.82	45.73	21.30	40.66

The database for this testing was the “863” dictation testing corpus which consists of 240 utterances, 3,145 syllables spoken by six male speakers. Our basic acoustic units were 138 consonant and vowel models (Ma, et al. 1996). The baseline syllable recognition accuracy was 78.1%. When the bigram language model was applied, the character accuracy of the baseline system was 90.1%. The tone model generated by the decision tree was used in this test. After integrating the tone information into the search process, the final recognition result improved to 91.7% with a character error rate reduction of 16.2%

7. Conclusion

In this paper, a Stochastic Polynomial Tone Model (SPTM) was proposed for tone modeling. Compared with the traditional HMM tone model, it achieved a remarkable error rate reduction for tone recognition. In order to further enhance model characterization, we proposed the Mixture Stochastic Polynomial Tone Model (MSPTM). Experimental result showed it can further improve tone recognition performance. A decision tree based clustering method was also employed to learn the tone pattern variations in continuous speech. A total of 28 tone patterns were obtained, and results showed that the effects of neighboring tones are the most important factors for tone pattern variation, and that Consonant/Vowel type of the syllable together with syllable position are also influential factors. A novel approach to integrate tone information into search processes was discussed. Experiments on continuous Mandarin speech recognition showed the character recognition error rate was reduced by 16.2%, which supported the effectiveness of our framework.

References

Cao, Y., Huang, T.-Y., Xu, B., and Li, C.-R. (2000). A stochastic polynomial tone model for continuous Mandarin speech. *ICSLP’2000 Proceedings*.

Chang, P.-C, Sun, S.-W., and Chen, S.-H. (1972). Mandarin Tone recognition by multilayer perception. *IEEE Trans. On Audio and Electroacoustic*, *20*:367–377.

Chen, C.J., Gopinath, R.A., and Monkowshi, M.D. (1997). New method in continuous Mandarin speech recognition. In *ICASSP’97 Proceedings (CDROM)*.

Chen, S.-H., Hwang, S.-H., and Wang, Y.-R. (1998). An RNN-based prosodic information synthesizer for Mandarin text-to-speech. *IEEE Trans. on Speech and Signal Processing*, *6*(3):226–239.

Chen, S.-H. and Wang, Y.-R. (1995). Tone recognition of continuous Mandarin speech based on neural networks. *IEEE Trans. on Speech and Signal Processing*, *3*(2):146–150.

Dempster, A.P., Larid, N.M., and Rubin, D.B. (1977). Maximum-likelihood from Incomplete Data via the EM algorithm. *Journal of Royal Statistical Society Series B*, *39*:13–18.

Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society (B)*, *58*:155–176.

Huang, H. and Seide, F. (2000). Pitch tracking and tone features for Mandarin speech recognition. *ICASSP’2000 Proceedings*, pp. 1523–1526.

Jain, A.K. et al. (2000). Statistical pattern recognition: A review. *IEEE Trans on Pattern Analysis and Machine Intelligence*, *22*(1):4–37.

Juang, B.H. and Katagiri, S. (1992). Discriminative learning for minimum error training. *IEEE Trans. on Signal Processing*, *40*(12):3043–3051.

Juang, B.H., Chou, W., and Lee, C.-H. (1997). Minimum classification error rate methods for speech recognition. *IEEE Trans. on Speech and Audio Processing*, *5*(3):257–265.

Lee, T., Carlson, R., and Granstorm, B. (1998). Context-dependent duration modeling for continuous speech recognition. *ICSLP’98 Proceedings (CDROM)*.

Lin, M.-C. (1998). *The Acoustic and Perceptual Characteristics of Chinese Mandarin Speech*. Chinese Language (in Chinese), No. 2.

Ma, B. et al. (1996). Context-dependent acoustic models in Chinese speech recognition. In *ICASSP’96 Proceedings (CDROM)*.

Russell, M. and Moore, R. (1985). Explicit modeling of state occupancy in Hidden Markov models for automatic speech recognition. *ICASSP'1985, Proceedings*, pp. 2376–2379.

Wang, C. and Seneff, S. (1998). A study of tone and tempo in continuous Mandarin digital strings and their application in telephone quality speech recognition. *ICSLP'98 Proceedings*, pp. 695–698.

Wang, H.-M. et al. (1997). Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data. *IEEE Trans. on Speech and Audio Processing*, *5*(2):196–201.

Wang, Y.R. et al. (1994). Tone recognition of continuous Chinese speech based on Hidden Markov model. *Int. J. Pattern Recognition and Artificial Intelligence*, *8*(1):233–246.

Wong, Y.W. and Chang, E. (2001). The effect of pitch and lexical tone on different Mandarin speech recognition tasks. *Eurospeech'2001 Proceedings (CDROM)*.

Zhao, L. et al. (1997). HMM based recognition of Chinese tones in continuous speech. The First China-Japan Workshop on Spoken Language Processing Proceedings (CDROM).