Acta Technica *62* (2017), No. 7A, 519–536 © 2017 Institute of Thermomechanics CAS, v.v.i.

Uyghur speech synthesis method based on hybrid primitive waveform splicing

Palidan Muhetaer[1], Wushouer Silamu[1], Maimaitiayifu[1]

Abstract. Aimed at not strong generalization performance existing in Uyghur speech synthesis system for traditional hard decision tree, one kind of binary soft decision tree algorithm was designed and improved to realize parameter estimation for Uyghur speech synthesis model based on contextual factors. Internal node selection was conducted according to child node membership, and context was distributed to several overlapped leaf nodes to improve model generalization and function approximation performance. The maximum entropy smoothing distribution was adopted to conduct feature capturing for local first moment and global second moment to realize the maximum likelihood estimation for soft decision parameter of HMM output probability distribution. Finally, through algorithm proposed in stimulation verification, under the premise of meeting real-time requirements for application, Uyghur speech synthesis effect could be effectively improved.

Key words. Decision tree, Generalization performance: Uyghur speech synthesis, Probability distribution, Hidden Markov, Global second moment.

1. Introduction

Speech processing procedure of Uyghur is one comprehensive discipline based on signal processing and linguistics[1], and it is widely used in production and life with each passing day. At the same time, requirements for practical application to Uyghur speech synthesis technology are higher and higher[2]. Adopting synthetic algorithm for (HMM) Uyghur speech parameter statistic of hidden Markov process to realize relatively excellent performance can be obtained and sound conversion is the most popular method at present. But problem for influences of slightly smoothing synthesis effect, lacking detail and relatively low naturalness etc. existing in HMM Uyghur speech synthesis on tone quality shall be solved[3].

[1]College of Information Science and Engineering, Xinjiang University, Urumqi, 830046, China

Specific procedure for parameter estimation of HMM Uyghur speech synthesis is [4∼5]: firstly, based on linguistics and grammatical rule, synthesis context information needed shall be obtained, and shall be labeled in synthesis model label; secondly, decision tree shall be obtained through training for Uyghur speech label that needs to be synthesized, and similar leaf node in HMM context model can be provided with decision then. Thirdly, according to model parameter obtained from decision, frequency spectrum and fundamental frequency parameter shall be further synthesized. Data frames in different quantities and states shall be obtained by taking advantage of duration model, and multidimensional parameter values for data frame of corresponding state duration shall be obtained according to variance and mean value of frequency spectrum and fundamental frequency model. Based on dynamic Uyghur speech feature, Uyghur speech synthesis parameter shall be provided with final estimation. Fourthly, according to solved parameter, Uyghur speech origin smoothing shall be established to synthesize Uyghur speech[6-10].

In algorithm research combined with decision tree, Literature [11] is one kind of clustering algorithm of hard decision tree in essence. In HMM clustering algorithm of traditional hard decision tree, based on HMM Uyghur speech synthesis and conversion, binary system (that is hidden fission process for decision tree) is solved by utilizing multi-variant contextual factors. Every model parameter is distributed to single leaf mode, and this kind of "divide and rule" method will cause data sparseness and poor generalization performance. Unobservable contextual parameters can not be accurately predicted, and its essence is one kind of weak function approximation. In order to solve this problem, Uyghur speech synthesis algorithm for the maximum likelihood for soft decision tree of the maximum entropy of hidden Markov was proposed here. Internal node selection was conducted according child node membership, and context was distributed to several overlapped leaf nodes to improve model generalization and function approximation performance. The maximum entropy smoothing distribution was adopted to conduct feature capturing for local first moment and global second moment to realize the maximum likelihood estimation for soft decision parameter of HMM output probability distribution.

2. F0 model of hard decision tree

2.1. F0 model under HMM framework

Fundamental frequency and its derivative and second derivative composition rely on three data streams for multi-space probability distribution of context from the left to the right. Acoustic unit trajectory is generated by observant value released by hidden state for this model. Output distribution of state relies on multi-space Gaussian distribution of context, and related contexts are assembled into groups by using decision tree to reduce parameter number. Visualization for context modeling is allowed. In order to simply express, the following discussion is only subject to HMM with signal data stream, which is simpler for multiple data streams.

Equivalent dynamic Bayesian network (DBN) used for HMM is given in Fig.1. In the diagram, q_t, o_t and g_t respectively expresses state index, sound feature vec-

tor and spatial index at the time of t. When MDS two spaces are used to define output distribution, observant value of spatial index and Uyghur speech label are consistent. Last frame index t_j for contextual factor c_j, duration d_j and state j are also introduced in the diagram, and it is obviously $d_j = t_j - t_{j-1}$. It shall be noted that state periphery is latent variable, and it shall be provided with unsupervised training by utilizing expectation-maximization (EM).

It can be known from Fig.1 that HMM can be simplified through three distributions: firstly, it is probability distribution state $p_j\left(d_j \left| c_j \right.\right)$ for duration; secondly, it is pronunciation (spatial) probability distribution $\omega_j\left(g_t \left| c_j \right.\right)$; thirdly, it is output probability distribution $b_j\left(o_t \left| g_t, c_j \right.\right)$ for given pronunciation label. By taking advantage of these basic distributions, model shown in Fig.2 shall be considered, and observation likelihood (o, g, c) for given Uyghur speech can be decomposed into:

$$\begin{aligned} p\left(o, g|c; \lambda\right) &= \sum\nolimits_{t_1, t_2, \cdots, t_J} \prod\nolimits_{j=1}^{J} p_j\left(d_j|c_j\right) \\ &= \prod\nolimits_{j=1}^{J} \omega_j\left(g_t|c_j\right) b_j\left(o_t \left| g_t, c_j \right.\right). \end{aligned} \tag{1}$$

Where, J and λ respectively expresses the sum of state and model parameter.

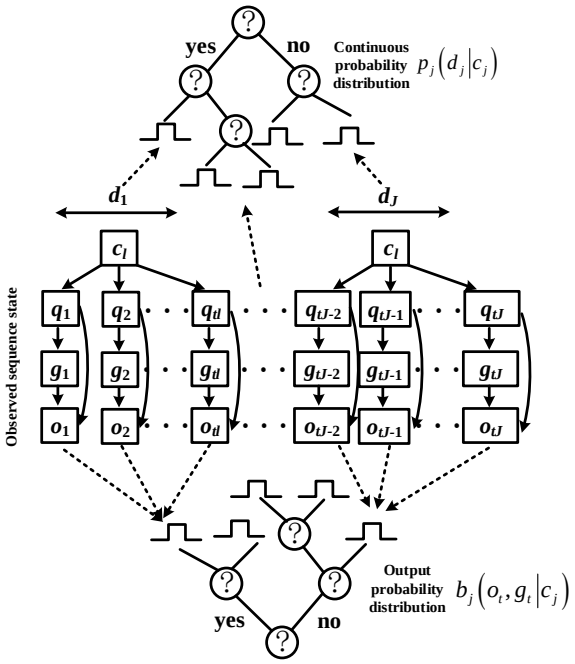


Fig. 1. HMM graph model

Given g_t is binary parameter: "1" expresses sound data frame, "0" expresses silent region. At the same time, supposing that b_j and p_j respectively are expressed through Gaussian distribution. Therefore, the above Uyghur speech likelihood func-

tion can be rewritten into:

$$\begin{aligned} p\left(o,g|c;\lambda\right) &= \sum\nolimits_{t_1,t_2,\cdots,t_J} \prod\nolimits_{j=1}^{J} \mathcal{N}\left(d_j;\bar{m}_j,\bar{\sigma}_j^2\right) \\ &= \prod_{t=t_{j-1}}^{t_j} \left[g_t\bar{\omega}_j\mathcal{N}\left(o_j;\bar{\mu}_j,\bar{\sum}_j\right)+\left(1-g_t\right)\left(1-\bar{\omega}_j\right)\right]. \end{aligned} \tag{2}$$

Where, $\mathcal{N}\left(\cdot;\mu,\sum\right)$ expresses Gaussian distribution of which mean vector is μ and variance matrix is $\sum$. In the Equation, duration and output distribution are through mean value $\bar{m}_j$ for duration, time variance $\bar{\sigma}_j^2$, voiced degree $\bar{\omega}_j$, output mean vector $\bar{\mu}_j$ and observant covariance matrix $\bar{\sum}_j$. Just like mentioned above, typical decision tree structure can be used to express basic distribution. Supposing $I_l^d\left(c_j\right)$ and $I_l^o\left(c_j\right)$ are defined as decision tree functions for binary index of output distribution and duration, of which l and c_j respectively expresses leaf index and contextual factor of state j, that is $I_l^d\left(c_j\right)$ and $I_l^o\left(c_j\right)$ can be used to confirm whether state j is distributed to the l duration and observant decision tree. By utilizing these index functions of decision tree, model parameter for hidden Markov can be expressed:

$$\begin{cases} m_j = \sum_l I_l^d(c_j)m_l, \sigma_j^2 = \sum_l I_l^d(c_j)\sigma_l^2 \\ w_j = \sum_l I_l^0(c_j)w_l, \mu_j = \sum_l I_l^0(c_j)\mu_l \\ z_j = \sum_l I_l^0(c_j)Z_l, \end{cases} \tag{3}$$

Where, m_l and σ_l^2 are respectively located in mean value and variance value for duration on the l leaf in time decision tree. ω_l, μ_l and $\sum_l$ respectively express probability distribution parameter of Uyghur speech expression and output to train the l leaf of output decision tree.

2.2. Parameter estimation for hidden Markov model

The maximum likelihood criterion is often used to estimate HMM model parameter. However, state periphery has hidden Uyghur nature. Therefore, EM algorithm shall be adopted to estimate. Given N Uyghur speeches $\{(o^n,g^n)\}_{n=1}^N$ of independent identical distribution, accompanying its corresponding contextual factor $\{c^n\}_{n=1}^N$, EM algorithm can be used to obtain the following parameter estimation equation:

$$\hat{m}_l = \frac{\sum_{n=1}^N \sum_{j=1}^{J^n} I_l^d(c_j^n) \sum_{t_j,t_j-1} \chi_j^n(t_j,t_{j-1})[t_j - t_{j-1}]}{\sum_{n=1}^N \sum_{j=1}^{J^n} I_l^d(c_j^n) \sum_{t_j,t_j-1} \chi_j^n\left(t_j,t_{j-1}\right)}. \tag{4}$$

$$\hat{\sigma}_l^2 = \frac{\sum_{n=1}^N \sum_{j=1}^{n} I_l^d(c_j^n) \sum_{t_j,t_j-1} x_j^n(t_j,t_{j-1})[t_j - t_{j-1} - \hat{m}_l]^2}{\sum_{n=1}^N \sum_{j=1}^{J^n} I_l^d(c_j^n) \sum_{t_j,t_j-1} x_j^n\left(t_j,t_{j-1}\right)}. \tag{5}$$

$$\hat{\mu}_l = \frac{\sum_{n=1}^{N}\sum_{j=1}^{n} I_l^0(c_j^n)\sum_t \gamma_j^n(t) g_t^n[o_t^n]}{\sum_{n=1}^{N}\sum_{j=1}^{J^n} I_l^0\left(c_j^n\right)\sum_t \gamma_j^n(t)} . \tag{6}$$

$$\hat{\sum}_l = \frac{\sum_{n=1}^{N}\sum_{j=1}^{n} I_l^0(c_j^n)\sum_t \gamma_j^n(t) g_t^n[(o_t^n-\hat{\mu}_l)(o_t^n-\hat{\mu}_l)^T]}{\sum_{n=1}^{N}\sum_{j=1}^{n} I_l^0(c_j^n)\sum_t \gamma_j^n(t) g_t^n} . \tag{7}$$

$$\hat{\omega}_l = \frac{\sum_{n=1}^{N}\sum_{j=1}^{J^n} I_l^0(c_j^n)\sum_t \gamma_j^n(t) g_t^n}{\sum_{n=1}^{N}\sum_{j=1}^{J^n} I_l^0(c_j^n)\sum_t \gamma_j^n(t)} . \tag{8}$$

Where, during the process of EM algorithm implementation, $\hat{m}_l\hat{\sigma}_l^2\hat{\mu}_l\hat{\sum}_l$ and $\hat{\omega}_l$ are respectively update values for $m_l\sigma_l^2\mu_l\sum_l$ and ω_l. At the same time, $\chi_j\left(t_j, t_{j-1}\right)$ is probability of state j from time t_{j-1} tot_j. $\gamma_j\left(t\right)$ expresses posterior probability of state j at the time of t. These probabilities can be calculated through famous forward-backward algorithm.

2.3. Uyghur speech state clustering of decision tree

Typical decision tree can be designed under HMM framework in general. Decision tree conducts structure through greedy and top-down iterative program, and log likelihood criterion can be improved to the maximum. This process starts from single root node, and expression for all Uyghur speech segments can be realized. During per iteration, the optimal terminal node shall be selected to make split terminal node present the maximum log likelihood increase to selected problem result. Fission process continues until it meets the termination criterion (such as the maximum description length (MDL) criterion). Overall log likelihood increase $\delta\mathcal{L}$ can be realized through two child nodes for l_2 and l_3 of split parent node l_1, and can be calculated through the following equation:

$$\begin{aligned}\delta\mathcal{L} =& \frac{1}{2}\log\left(\left|\hat{\sum}_{l_1}\right|\right)\sum_{n=1}^{N}\sum_{j=1}^{J^n} I_{l_1}^0(c_j^n)\sum_t \gamma_j^n(t) - \\ & \sum_{l\in\{l_2,l_3\}}\frac{1}{2}1og\left(\left|\hat{\sum}_{l_1}\right|\right)\sum_{n=1}^{N}\sum_{j=1}^{J^n}(c_j^n)\sum_t \gamma_j^n(t) .\end{aligned} \tag{9}$$

Where, superscript n is training sample quantity for Uyghur speech. It shall be noted that assumption shall be given in order to obtain likelihood probability increase: (1) possession probability value is constant during the process of clustering; (2) it is supposed that overall likelihood measure is subject to one simple posterior probability weighting of log likelihood for mean approximation. These assumptions make $\delta\mathcal{L}$ calculation of terminal node possible.

3. HMM clustering of soft background

Decision tree is a hierarchical structure containing internal node and terminal leaf. Every terminal node can capture clustering statistics features in linguistic text. For given context c, every internal node shall be provided with binary test $f_m(c)$ as well. One test child node shall be selected according to test result. Supposing that $I_m(c)$ is defined as indicator function of node m. Supposing that $I_{m_L}(c)$ and $I_{m_R}(c)$ express indicator functions for child node on its left and right as well; $I_{m_L}(c)$ and $I_{m_R}(c)$ can be calculated as:

$$I_{m_L}(c) \stackrel{\text{det}}{=} \begin{cases} I_m(c), \text{ if } f_m(c) = true \\ 0, \text{ if } f_m(c) = false \end{cases} \tag{10}$$

$$I_{m_R}(c) \stackrel{\text{det}}{=} \begin{cases} I_m(c), \text{ if } f_m(c) = false \\ 0, \text{ if } f_m(c) = true \end{cases} \tag{11}$$

Therefore, in order to confirm given factor distribution in linguistic context, it shall be started from node to recursively apply it in per internal node test, and one branch result shall be selected according to output. Therefore, there is only one path from root node to terminal node for hard decision tree, and Uyghur speech segment shall be distributed here to affect distribution of single leaf. In order to improve typical decision tree performance, soft binary decision tree structure was proposed here to be able to establish multiple fuzzy paths from root to multiple leaves.

3.1. Algorithm structure

All offspring individuals are provided with redirection by soft decision tree by taking advantage of soft-decision $\bar{f}_m(c)$ in its internal node, but it has some membership. It can be calculated according to $\bar{f}_m(c)$ and 1-$\bar{f}_m(c)$. In fact, soft decision tree for per node represents fuzzy subset of boundary factor space. Therefore, every context is attached to several nodes. More accurately, when given context c is provided with node m traversal by us, soft inquiry $\bar{f}_m(c)$ represents sub-generational membership level on the left. Obviously, 1-$\bar{f}_m(c)$ represents sub-generational membership level on the right.

In hard and soft decision tree based on HMM, firstly, one set of contextual factor shall be defined, and all Uyghur speech training samples shall be extracted. Then, relative to hard decision tree inquire $f_m(c)$, a large quantity of soft inquire problems (soft test) $\bar{f}_m(c)$ is designed for different contextual factors. These problems are finally distributed to internal node of decision tree to make sub-generational fuzzy decision rather than final weak decision shall be made.

Based on the above discussion, all terminal leaves may be effective to any context. At the same time, it is necessary to express indicator function $I_m(c)$ into membership function form shown in Equation (3) by taking advantage of context c and node m.

Membership function $\bar{I}_m(c)$ can be calculated according to the following equation:

$$\left\{ \begin{array}{l} Initialization: \ I_{root}(c) = 1 \\ Recursion: \left\{ \begin{array}{l} \bar{I}_{m_L}(c) = \bar{f}_m(c)\bar{I}_m(c) \\ \bar{I}_{m_R}(c) = (1 - \bar{f}_m(c))\bar{I}_m(c) \end{array} \right\} \end{array} \right. \tag{12}$$

Where, m_L and m_R are respectively child nodes on the left and right of node m. According to recursive degree defined in the above, all memberships can be calculated according to traverse tree mode. Traverse tree starts from that root node membership is set as 1, and child node on its left and right can be obtained after confirming its membership $\bar{I}_m(j)$ through observing node m. In case it is child node on the left, its membership can be calculated through $\bar{f}_m(c)\,\bar{I}_m(c)$; otherwise, program will return to $\left(1 - \bar{f}_m(c)\right)\bar{I}_m(c)$, of which m is parent node.

In training stage, soft decision $\bar{f}_m(c)$ conducts selection through predefined contextual function. The following conditions shall met for this function to all contextual factors:

$$\forall m, c, 0 \leq \bar{f}_m(c) \leq 1\,. \tag{13}$$

During the process of defining soft problem, the above constraints shall be considered. That is circumstance that soft problem value is more than 1 or less than 0 is not allowed to exist. Therefore, before decision tree starts to cluster, these kinds of problems shall be provided with normalization steps.

3.2. HMM clustering distribution of soft context for the maximum entropy

Proposed HMM clustering mode for soft context and HMM clustering mode for hard decision tree are subject to the same structural model. Therefore, model likelihood expression shown in Equation (1) is still effective to proposed model. What is different is that there is difference in capturing mode for fixed context relying relation on F0 track. In addition, expression method for output distribution $b_j(\cdot)$ in Equation (1) is different, and the maximum entropy model (MEM) was adopted here to guarantee distribution estimation stability.

The maximum entropy principle is to improve target entropy (uncertainty) furthest according to observant vector knowledge, and it is called effective estimation. For the maximum output distribution model $b_j(o_t\,|g_t, c_j)$ of given Uyghur speech label, supposing that trained Uyghur speech includes Uyghur speech label for independent identical distribution, D-dimension output feature vector $\{o_t\}_{t=1}^{T}$ may be affected by some contextual information $\{c_t\}_{t=1}^{T}$. Contextual information clusters through membership function $\{I_l(\cdot)\}_{l=1}^{L}$ of soft decision tree structure. In terms of the maximum entropy principle, one set of constraint distribution shall be specified, and one distribution approximate to average shall be selected through optimization entropy criterion. In fact, this kind of scheme is one kind of distribution modeling mode having preference.

$$b(o|g,c) \stackrel{\text{det}}{=} \arg\max_{b} \mathcal{H}(b(o|g,c)\}\,. \tag{14}$$

Where, $\mathcal{H}$ is entropy measurement, and it can be defined as:

$$\mathcal{H}(b(o|g,c)\} \stackrel{\text{det}}{=} -\sum_{t=1}^{T}\int b(o|g_t,c_t)1ogb(o|g_t,c_t)do\,. \tag{15}$$

Considering the following constraints:

$$\begin{cases} \forall 1 \le l \le L, \forall \mathrm{c}, \mathrm{g}\int\limits_{o} \mathrm{b}(\mathrm{o}|\mathrm{g},\mathrm{c}) = 1 \\ E\left\{goo^T\right\} = \bar{E}\left\{goo^T\right\} \\ E\left\{I_l\left(c\right)go\right\} = \bar{E}\left\{\bar{I}_l\left(c\right)go\right\} \end{cases} \tag{16}$$

The first constraint shall guarantee that the total distribution is 1. At the same time, E and $\bar{E}$ indicate that real mathematical expectation can be given through the following equation:

$$E\{\bar{I}_l\left(c\right)go\} = \sum_{t=1}^{T}\bar{I}_l(c_t)g_t\int\limits_{o} ob(o|g_t,c_t)do\,. \tag{17}$$

$$E\{\bar{I}_l\left(c\right)go\} = \sum_{t=1}^{T}\bar{I}_l(c_t)g_to_t\,. \tag{18}$$

$$E\{goo^T\} = \sum_{t=1}^{T}g_t\int\limits_{o} oo^T(o|g_t,c_t)do\,. \tag{19}$$

$$E\{goo^T\} = \sum_{t=1}^{T}g_to_to_t^T\,. \tag{20}$$

These constraints make that estimation distribution can capture partial first moment $\bar{E}\left\{\bar{I}_l\left(c\right)go\right\}$ of trained Uyghur speech data, and global secondary moment $\bar{E}\left\{goo^T\right\}$. In order to solve equation constraint optimization, Lagrange multiplier method can be selected:

$$b(o|g,c) \stackrel{\text{det}}{=} \arg\max_{b}\mathcal{J}\left(b\right)\,. \tag{21}$$

$$\begin{aligned} \mathcal{J}\left(b\right) =& \mathcal{H}(b(o|g,c)) + \lambda_{bo}\left[\int_{o} b(o|g,c)do\right] \\ &+ \sum_{l=1}^{L}\lambda_{bl}^T\left[E\left\{\bar{I}_l\left(c\right)go\right\} - \bar{E}\left\{\bar{I}_l\left(c\right)go\right\}\right] \\ &+ \left[E\left\{go^T \wedge o\right\} - \bar{E}\left\{go^T \wedge o\right\}\right]\,. \end{aligned} \tag{22}$$

Where, $\mathcal{J}\left(b\right)$ expresses new optimization function; λ_{b0}, λ_{b1} and v are used to

eliminate Lagrange multiplier of equation constraint. $\mathcal{J}(b)$ shall be provided with derivation by taking advantage of output probability distribution $b(o|g,c)$, and it shall be set as zero. Equation can be obtained:

$$\begin{aligned}\frac{\partial \mathcal{J}(b)}{\partial b} = &\sum_{t=1}^{T}\int_{o}[-1ogb(o|g_t,c_t)-1+\lambda_{b0}\\ &+g_t o^T\wedge o+\sum_{l=1}^{L}\lambda_{bl}^T\bar{I}_l(c_t)\,g_t o=0\,.\end{aligned} \tag{23}$$

One obvious solution that can meet the above equation is:

$$1ogb(o|g_t,c_t)=g_t o^T\wedge o+\sum_{l=1}^{L}\lambda_{bl}^T\bar{I}_l(c_t)\,g_t o-1+\lambda_{b0}\,. \tag{24}$$

Therefore, $b(o|g_t,c_t)$ is one simple Gaussian distribution, and it can be expressed as:

$$b(o|g,c)=\mathcal{N}\left(o;\sum_{l=1}^{L}\bar{I}_l(c_t)\,\mu_l,\sum\right)\,. \tag{25}$$

Where, $\mathcal{N}$ is Gaussian distribution; μ_l is mean vector for l leaf parameter of decision tree; $\sum$ is $D\times D$ covariance matrix used for all leaves.

3.3. Parameter estimation

On the basis of HMM clustering structure for soft environment in the last section, its parameter estimation method is discussed here. In training stage, one set of N trained Uyghur voice samples for independent identical distribution containing acoustic feature $\{o^n\}_{n=1}^N$, Uyghur speech label $\{g^n\}_{n=1}^N$ and contextual factor $\{c^n\}_{n=1}^N$. Objective is to find the optimal model parameter $\hat{\lambda}$ to make likelihood measurement maximized.

$$\begin{cases}\hat{\lambda}\overset{\text{det}}{=}\arg\max_{\lambda}\mathcal{L}(\lambda)\\ \mathcal{L}(\lambda)\overset{\text{det}}{=}\sum\limits_{n=1}^{N}\ln p\left(o^n,g^n\,|c^n\,;\lambda\right)\end{cases} \tag{26}$$

In this section, supposing that soft decision tree structure has been trained, it is just needed to find maximum likelihood estimation of its parameter λ. In the next section, how to train the best soft decision tree structure will be described. Similar to typical HMM model, the maximum likelihood expression of Equation (1) can cause very complex optimization, and it seems that it is impossible to directly get solution. The main problem is distribution is latent variable determined by boundary of state. EM technology provides iterative algorithm that can conquer this problem, and $\hat{\lambda}$

can be obtained through iterating the following function to the maximum:

$$\lambda^{r+1} = \arg\max_{\lambda} \mathcal{Q}\left(\lambda; \lambda^{r}\right) . \tag{27}$$

$$\begin{aligned} \mathcal{Q}\left(\lambda; \lambda^{r}\right) = & \sum_{n} [\sum_{t_j, t_{j-1}} \chi_j^n \left(t_j, t_{j-1}; \lambda^{r}\right) \log p_j \left(t_j - t_{j-1} \left| c_j^n \right.\right) \\ & + \sum_{t} \sum_{j} \gamma_j^n \left(t; \lambda^{r}\right) \left\{ \log \omega_j \left(g_t^n \left| c_j^n \right.\right) + \log b_j \left(o_t^n \left| g_t^n, c_j^n \right.\right) \right\} . \end{aligned} \tag{28}$$

Where, χ_j and γ_j are occupation probabilities of the above definition, and r is EM iterative label. n is No. range of Uyghur speech. In order to estimate the best setting of parameter, partial derivative Q value of all model parameters λ can be taken as 0. Calculation processes for these partial derivatives are shown in the following:

$$\begin{aligned} \frac{\partial \mathcal{Q}\left(\lambda; \lambda^{r}\right)}{\partial \mu_l} = & \sum^{-1} \sum_{n} \sum_{t} \sum_{j} \gamma_j^n \left(t; \lambda^{r}\right) \\ & \bar{I}_l(c_j^n) \left(o_t^n - \sum_{l=1}^{L} \bar{I}_l(c_j^n) \mu_l \right) . \end{aligned} \tag{29}$$

$$\begin{aligned} \frac{\partial \mathcal{Q}\left(\lambda; \lambda^{r}\right)}{\partial \sum} = & \frac{1}{2} \sum\nolimits^{-1} \sum_{n} \sum_{t} \sum_{j} \gamma_j^n \left(t; \lambda^{r}\right) \\ & \left\{ -1 + \left(o_t^n - \sum_{l=1}^{L} \bar{I}_l(c_j^n) \mu_l \right)^{T} \left(o_t^n - \sum_{l=1}^{L} \bar{I}_l(c_j^n) \mu_l \right) \sum^{-1} \right\} . \end{aligned} \tag{30}$$

Supposing the Equation (30) is zero, the maximum likelihood estimation of its model parameter can be obtained. Equation set shall be set, and the optimal vector $\{\bar{\mu}_l\}_{l=1}^{L}$ for mean parameter can be calculated:

$$R\hat{\mu} = P . \tag{31}$$

Where, $\hat{\mu}$ is $L \times D$ matrix including mean parameter. R and P are respectively $L \times L$ and $L \times D$ matrix, and they can be defined as:

$$\begin{cases} R = [r_{uv}]_{L \times L} , r_{uv} = \sum_{n} \sum_{j} \bar{I}_u \left(c_j^n\right) \bar{I}_v \left(c_j^n\right) \sum_{t} \gamma_j^n \left(t, ; \lambda^{r}\right) \\ P = [p_u]_{L \times L} , p_u = \sum_{n} \sum_{j} \bar{I}_u \left(c_j^n\right) \bar{I}_v \left(c_j^n\right) \sum_{t} o_t^n \gamma_j^n \left(t, ; \lambda^{r}\right) \end{cases} \tag{32}$$

Through setting that partial derivative of covariance matrix $\sum$ is 0, covariance

matrix $\sum$ shall be calculated:

$$\bar{\sum} = \left(o_t^n - \sum_{l=1}^{L} \bar{I}_l(c_j^n)\mu_l \right)^T \left(o_t^n - \sum_{l=1}^{L} \bar{I}_l(c_j^n)\mu_l \right) . \tag{33}$$

Output probability distribution for soft decision tree is obtained through simple procedure training in the above, In order to realize soft decision tree clustering, log likelihood measurement for the optimal model parameter needs to be calculated:

$$\mathcal{L} \propto -\frac{1}{2} 1og \left(\left| \hat{\sum} \right| \right) \sum_n \sum_t \sum_j \gamma_j^n (t) . \tag{34}$$

Where, $|\cdot|$ represents operator of matrix determinant.

3.4. Fuzzy soft decision clustering

In order to automatically capture relying relation between acoustic feature and contextual factor, one kind of construction algorithm for soft decision tree was proposed. Similar to construction process for traditional hard decision tree, through greedy and top-down program, log likelihood measurement of soft decision tree can be improved at the greatest extent. Advantage of traditional hard decision tree is that its terminal node that can be divided independently. In hard decision tree, terminal node represents non-overlapping domain in contextual factor space; therefore, after splitting of leaf node, changes in obtained objective values are effective to other nodes as well, and double counting is not needed. But during the process of fuzzy soft decision clustering, it does not meet any more. Therefore, relative to traditional hard decision tree structure, split process for all terminal nodes shall be calculated, and calculated amount increases. Typical structure for Uyghur speech synthesis system is shown in Fig.2. Construction process for soft decision tree is shown in pseudo-code 1.

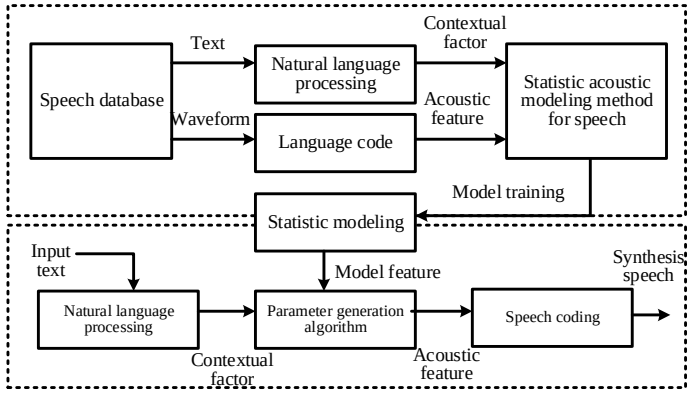


Fig. 2. Structure for uyghur speech synthesis system

The main difference with hard decision tree process is different evaluation quantity of conducting clustering process during the process of per iteration. In hard clustering, two new leaf nodes are just needed to be evaluated, but all leaf nodes need to be assessed in soft clustering, which increases soft clustering computing complexity of an order of magnitude. Supposing that decision tree having L leaf node shall be constructed, Q inquires shall be defined. $(2L-3)\,Q$ is needed to calculate complexity for hard decision tree clustering during this process, but soft decision tree clustering needs $[L(L-1)/2]Q$ complexity.

Pseudo-code 1.
Input: contextual factor c_j^n; observant feature o_t^n; state possession probability $\gamma_j^n(t)$; predefined soft decision $\bar{f}_m(c)$.
Output: the final node amount of decision tree$L$; membership function $\left\{\bar{I}_l(c)\right\}_{l=1}^{L}$ for context relying of terminal node for decision tree.
Soft context clustering algorithm:
Initialization $L=1$ and $f_1(c)=1$;
While termination condition does not meet do
For $l=1,2,\cdots,L$ (per node) do
$\forall j,n\ \bar{I}_{L+1}\left(c_j^n\right)=\left(1-\bar{f}_m\left(c_j^n\right)\right)\bar{I}_l\left(c_j^n\right)$;
$\forall j,n\ \bar{I}_l\left(c_j^n\right)=\bar{f}_m\left(c_j^n\right)\bar{I}_l\left(c_j^n\right)$
Calculate R and P according to Equation (32);
Calculate $\hat{\mu}$ according to Equation (31);
Calculate $\bar{\sum}$ according to Equation (33);
Calculate $\mathcal{L}(l,m)$ according to Equation (34);
Endfor
$\hat{l},\hat{m}=\arg\max_{l,m}\mathcal{L}(l,m)$;
$\forall j,n\ \bar{I}_{L+1}\left(c_j^n\right)=\bar{f}_{\hat{m}}\left(c_j^n\right)\bar{I}_{\hat{l}}\left(c_j^n\right)$;
$\forall j,n\ \bar{I}_{\hat{l}}\left(c_j^n\right)=\bar{f}_{\hat{m}}\left(c_j^n\right)\bar{I}_{\hat{l}}\left(c_j^n\right)$
Endwhile

---

4. Experimental analysis

Uyghur speech database used in the experiment is called Nick, and the database is composed of Uyghur speeches for about 2500 Western Uyghur male. The database is collected by Xinjiang Laboratory of Multi-Language Information Technology to research Uyghur speech synthesis, and sentence length range is from 3 to 36 words. Average length is 7.3 words. In addition, the database covers the most common English words, and sentences and syllables combined with double phonemes, with 2944 different words. Sampling shall be conducted by taking advantage of Blackman window for 25 milliseconds of 5 milliseconds displacement at the place of 48 kHz for Uyghur speech waveform. Hardware configuration: CPU i7-5400,RAM4G ddr3-1600,win7 flagship. F0 modeling method proposed shall be provided with objective and subjective test. Contrast algorithm is subject to Literature [11,13], and these two kinds of algorithm are deformation for structural algorithm of hard decision tree.

4.1. Objective evaluation

Learning curve during the construction process for hard and soft decision tree is given in Fig.3, and 800 Uyghur speech examples shall be adopted to train. 400 Uyghur speech example shall be tested. Normalized log likelihood measure index described in Fig.1 can be calculated according to the following equation:

$$\mathcal{L} = \frac{1}{\sum_t \sum_l g_{tl}} \sum_t \sum_l g_{tl} \log b\left(o_{tl} \left| g_{tl} \right.\right). \tag{35}$$

Where, F0 derivative can be expressed as o_{tl}, and its Uyghur speech label can be expressed in g_{tl}. t is data frame index of Uyghur speech, and l expresses unequal dynamic or static feature from 1 to 3.

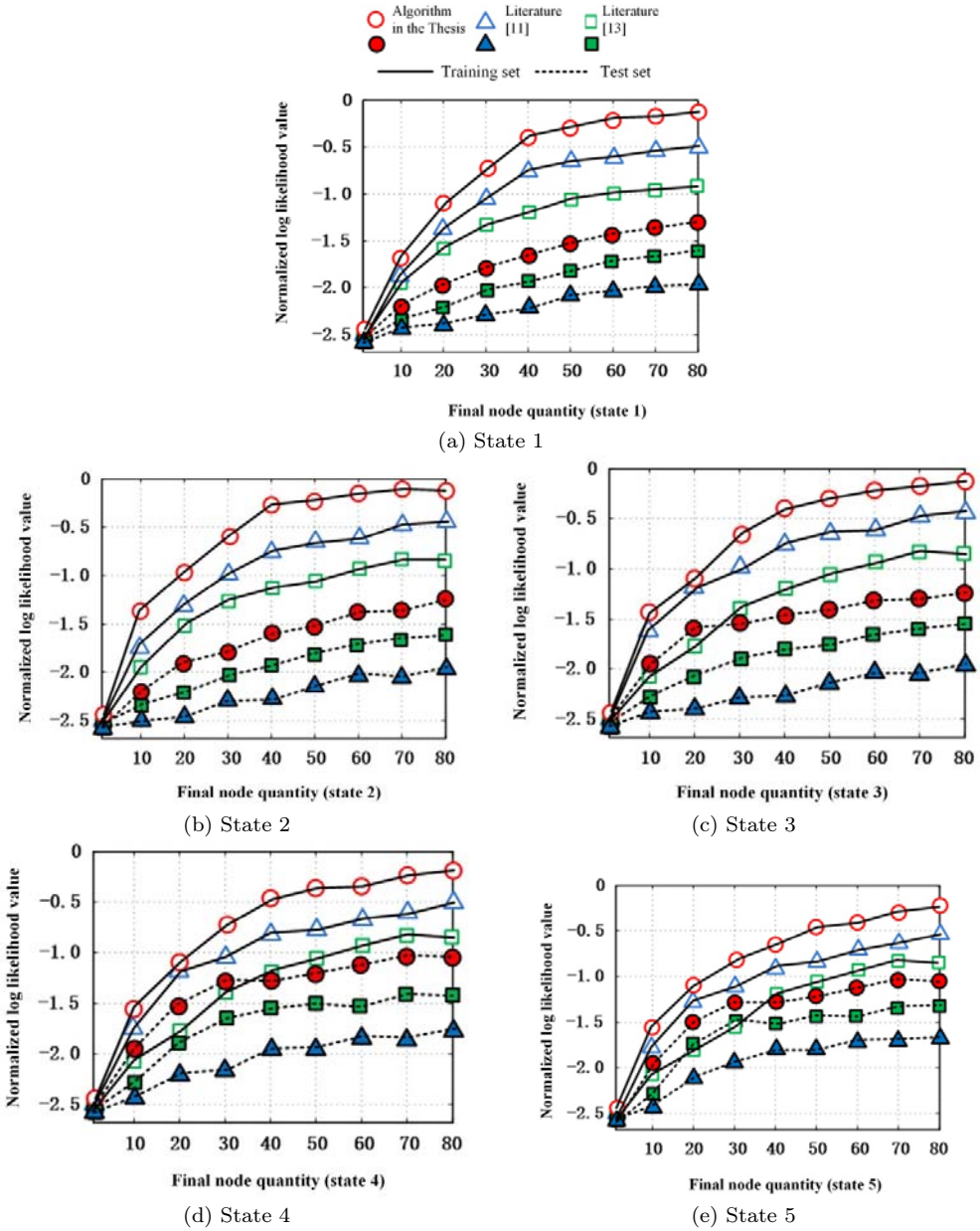
In Fig.3, test and training data shall be measured by using normalized log likelihood measure. Full line is normalized log likelihood measure for training data set, and hidden line expresses normalized log likelihood measure of test data set. At the same time, the final optimal leaf quantity shall be calculated and confirmed in diagram according to MDL principle. It can be known from experimental curve in Fig.3 that SDT-HMM algorithm is higher than algorithm in Literature [11] and Literature [13]. Therefore, better log likelihood measure can be realized under relatively small model parameter in SDT-HMM algorithm, and this process can compensate calculation complexity increase caused by soft decision tree described in 2.4 section. It can be knows from learning curve shown in Fig.3 that soft decision tree can provide better generalization ability compared with hard decision tree.

At the same time, Uyghur speech and root-mean-square error (RMSE) between F0 track of natural logarithm shall be adopted to conduct algorithm performance evaluation:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\sum_t g_t} \sum_t g_t \left(f_t^P - f_t^T\right)^2}. \tag{36}$$

Where, g_t, f_t^P and f_t^T are respectively Uyghur speech label, target logarithm F0 value and prediction logarithm F0 value of the t Uyghur speech frame. This measurement index used to calculate four training data set respectively, including Uyghur speech training set of 100,200,400,and 800 quantity. Experimental data is shown in Fig.4.

It can be known from Fig.4 that with the increase of sample quantity for Uyghur speech training, RMSE index of algorithm presents descending trend, which is consistent with the practical condition. At the same time, it can be known from the diagram that this kind of descending trend gradually becomes stable with the increase of training set and gradually approximate, which shows that advantage of proposed algorithm is more obvious under small training data set.



(a) State 1

(b) State 2

(c) State 3

(d) State 4

(e) State 5

Fig. 3. Normalized log likelihood value for HMM state considering leaf quantity

4.2. Subjective evaluation

Literature [14] is subject to 7 points to evaluate its mean value (-3∼3) to evaluate subjective similarity (CMOS) of synthesis and natural language. In CMOS test,

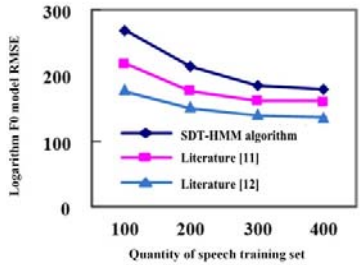


Fig. 4. RMSE index comparison data

according to audiences, better Uyghur language shall be selected, and difference between both shall be confirmed. For this, four levels shall be defined: 0, 1, 2 and 3, same, slight difference, difference and many differences. Evaluation comparison data based on CMOS are respectively shown in Fig.5 and Fig.6.

It can be seen from CMOS subjective evaluation value of context HMM in Fig.5 that with the increase of training data set, absolute value difference for evaluation between algorithms presents descending attitude. It can be seen from subjective evaluation for paired comparison experiment that with the increase of training data set, evaluation difference of SDT-HMM algorithm and contrast algorithm gradually descends. It is shown in the above experiment that SDT-HMM algorithm is more suitable for small data set circumstance, having higher practical value.

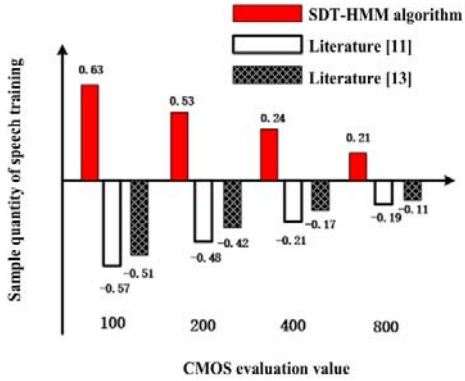
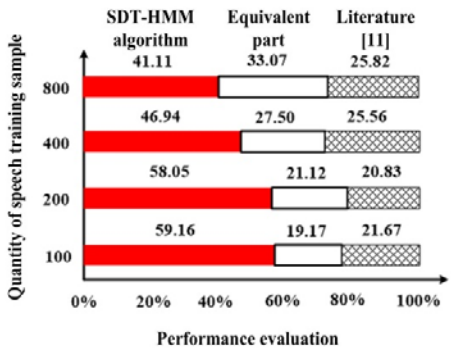


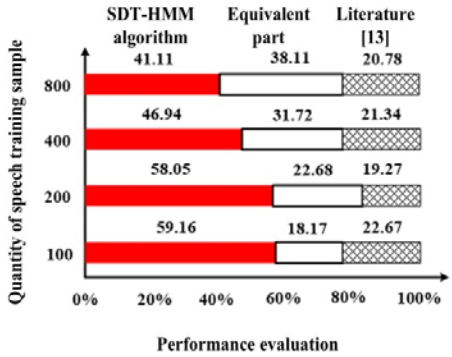
Fig. 5. CMOS subjective evaluation in context HMM

4.3. Calculation time of subjective evaluation

The above subjective evaluation experiment represents performance advantage of SDT-HMM algorithm, and calculation time index of algorithm is provided with experimental analysis in this section. 200-800 shall be selected for training sample quantity, and 100-400 shall be selected for test sample quantity. Selected database and hardware configuration are the same as the above. Algorithm training time, test time and the total time are shown in Table 1.



(a) Compare with Literature [11]



(b) Compare with Literature [13]

Fig. 6. Subjective evaluation for paired comparison experiment

It can be known from Table 1 that during the sample training process, although calculation complexity of SDT-HMM algorithm is higher than contrast algorithm, SDT-HMM algorithm can realize better log likelihood measure under the relatively small model parameter. The process can compensate calculation complexity increase caused by soft decision tree described in 2.4 Section. Therefore, calculation complexity of training process is relatively high. During the measurement link, for calculation complexity of SDT-HMM algorithm is relatively high, its calculation time is slightly higher then contrast algorithm. In general, the total calculation time of SDT-HMM algorithm is still less than that of contrast algorithm.

Table 1. Running time comparison of algorithm (s)

	Quantity	SDT-HMM	Literature [11]	Literature [13]
Training	200	13.2	15.7	16.2
	500	18.7	19.3	19.8
	800	22.3	23.4	24.1
Test	100	7.6	6.6	7.5
	250	9.3	8.9	8.7
	400	10.8	9.6	9.9

5. Conclusion

One kind of binary soft decision tree algorithm was proposed in the Thesis to realize parameter estimation for Uyghur speech synthesis model based on contextual factor to solve not strong generalization performance existing in Uyghur speech synthesis system of traditional hard decision tree. Effectiveness for proposed algorithm was verified in the experimental result. At the same time, it shall be noted that for fuzzy multiple path mode adopted by soft decision tree algorithm will cause calculation complexity of algorithm to correspondingly increase. Although in experimental link verification, SDT-HMM algorithm can realize better log likelihood measure under relatively small model parameter. This process can compensate calculation complexity increase. But how to lower calculation complexity is still one valuable research direction.

Acknowledgement

National "973" basic research project fund No.2014CB340506; Project supported by the National Natural Science Foundation of China No.U1603262.

References

[1] Y. Du, Y. Z. Chen, Y. Y. Zhuang, C. Zhu, F. J. Tang, J. Huang: *Probing Nanostrain via a Mechanically Designed Optical Fiber Interferometer.* IEEE Photonics Technology Letters, *29* (2017), 1348–1351.

[2] W. S. Pan, S. Z. Chen, Z. Y. Feng: *Automatic Clustering of Social Tag using Community Detection.* Applied Mathematics & Information Sciences, *7* (2013), No. 2, 675–681.

[3] A. Malousi, I. Chouvarda, V. Koutkias, et al.: *SpliceIT: A hybrid method for splice signal identification based on probabilistic and biological inference*[J]. Journal of Biomedical Informatics, *43* (2010), No. 2, 208–217.

[4] W. Peng, X. Zhang, Z. Gong Z., et al.: *Miniature Fiber-Optic Strain Sensor Based on a Hybrid Interferometric Structure*[J]. IEEE Photonics Technology Letters, *25* (2013), No. 24, 2385–2388.

[5] C. J. Zhang, H. M. Qiu, J. P. Qiu: *Relationship of polymorphisms in the cholesteryl ester transport protein gene R451Q with coronary heart disease and diabetes in Uyghur and Han Chinese.*[J]. Genetics & Molecular Research Gmr, *13* (2014), No. 1, 954–62.

[6] R. M. Liu, H. J. Liu, J. L. Cong, et al.: *Genetic characteristics of the couple with non-syndromic sensorineural hearing loss and fertility guidance.*[J]. International Journal of Clinical & Experimental Medicine, *8* (2015), No. 11, 21746.

[7] H. Zhang, M. Qiu, S. Kang S., et al.: *A tunable interferometric optical bandpass filter based on flexural acoustic wave modulation in a SMF-MOF hybrid configuration with fiber offset splicing output*[J]. Microwave & Optical Technology Letters, *57* (2015), No. 1, 31–35.

[8] Y. Y. Zhang, Q. Li, W. J. Welsh, P. V. Moghe, and K. E. Uhrich: *Micellar and Structural Stability of Nanoscale Amphiphilic Polymers: Implications for Anti-atherosclerotic Bioactivity*, Biomaterials, *84* (2016), 230–240.

[9] J. W. Chan, Y. Y. Zhang, and K. E. Uhrich: Amphiphilic Macromolecule Self-Assembled Monolayers Suppress Smooth Muscle Cell Proliferation, Bioconjugate Chemistry, *26* (2015), No. 7, 1359–1369.

[10] D. S. Abdelhamid, Y. Y. Zhang, D. R. Lewis, P. V. Moghe, W. J. Welsh, and K. E. Uhrich: *Tartaric Acid-based Amphiphilic Macromolecules with Ether Linkages Exhibit Enhanced Repression of Oxidized Low Density Lipoprotein Uptake*, Biomaterials, *53* (2015), 32–39.

[11] Y. Y. Zhang, A. Algburi, N. Wang, V. Kholodovych, D. O. Oh, M. Chikindas, and K. E. Uhrich: *Self-assembled Cationic Amphiphiles as Antimicrobial Peptides Mimics: Role of Hydrophobicity, Linkage Type, and Assembly State*, Nanomedicine: Nanotechnology, Biology and Medicine, *13* (2017), No. 2, 343–352.

[12] J. Liu J., H. Cui H., X. Dai X., et al.: *The total least squares method in multi-view point clouds splicing*[C]// International Congress on Image and Signal Processing. IEEE, (2010), 762–765.

[13] S. R. Hertz: *A model of the regularities underlying speaker variation: evidence from hybrid synthesis*[C]// INTERSPEECH 2006 - Icslp, Ninth International Conference on Spoken Language Processing, Pittsburgh, Pa, Usa, September. DBLP (2006).

[14] Y. Xing, P. Stoilov, K. Kapur, et al.: *MADS: a new and improved method for analysis of differential alternative splicing by exon-tiling microarrays.*[J]. Rna-a Publication of the Rna Society, *14* (2008), No. 8, 1470.

[15] E. Nolan, R. Filshie: *High efficiency transfection based on low electric field strength*, long pulse length: US, US6800484[P].

[16] M. U. Kismarton: (2010) Composite structural members and methods for forming the same: US, US 7721495 B2[P] (2004).

[17] L. D. Fielder: *Frame-based audio coding with video/audio data synchronization by dynamic audio frame alignment: Acoustical Society of America Journal*, US6124895[P] (2000).

Received May 7, 2017