A Hierarchical Viterbi Algorithm for Mandarin Hybrid Speech Synthesis System

Ran Zhang [1], *Zhengqi Wen* [1], *Jianhua Tao* [1], *Ya Li* [1], *Bing Liu*[1], *Xiaoyan Lou*[2]

[1] Institute of Automation Chinese Academy of Sciences, Beijing, China
[2] Samsung Telecom R&D Center, Beijing, China
rzhang@nlpr.ia.ac.cn, zqwen@nlpr.ia.ac.cn

Abstract

The hybrid speech synthesis system, which combines the hidden Markov model and unit selection method, has become an additional main stream in state-of-the-art TTS systems. However, traditional Viterbi algorithm is based on global minimization of a cost function and the procedure can end up selecting some poor-quality units with larger local errors, which can hardly be tolerated by the listeners. In Mandarin and many other languages, the naturalness of the region of consecutive voiced speech segments (CVS) is more essential to the overall quality of the synthetic speech. Consequently, in this paper, we proposed to use a hierarchical Viterbi algorithm which involves two rounds of Viterbi search: one is for the sub-paths in the CVS regions; the other is for the utterance path connecting all the sub-paths. In the proposed technique, we defined CVS Region as a region which is formed by two or more voiced phones, and whose observation of pitch has a continuous value. Subjective evaluations suggest that the use of hierarchical Viterbi algorithm in the Mandarin hybrid speech synthesis system outperforms the use of traditional algorithm in both the naturalness and speech quality of synthetic speech.

Index Terms: hierarchical Viterbi algorithm, unit selection, hidden Markov models, hybrid TTS, CVS Region

1. Introduction

Unit selection based waveform concatenation system [1, 2]and hidden Markov models (HMMs) based parametric system [3] are two mainstream systems of corpus-based TTS for decades. While conventional unit-selection based methods can preserve the naturalness of real speech, they often suffer from the occasional glitches or artifacts, especially with a small corpus. Worse still, the methods have a loose control on the speech features of a single unit, which usually leads to an inconsistent output speech. HMM-based synthesis methods are much more stable and have much littler footprint; however, the output speech carries an intrinsic hiss-buzz vocoding flavor due to their source-filter assumption, which leads to a poor performance on naturalness.

Now there is a growing trend to combine the advantages of these two systems together into one single hybrid system [4, 5]. In this hybrid system, the selection of the unit sequence is guided by the HMMs trained according to the criterion of Maximum Likelihood. On the one hand, the underlying HMM-based prediction can insure the smoothness and consistency of generated trajectories, which can guide unit selection to match several features such as spectrum, pitch and duration. On the other hand, the use of natural speech segments in concatenation preserves natural variation which is hard to model. Furthermore, the hybrid approach's cost function is usually made of the likelihood of sentence HMM, instead of the complex context based sub-costs, so that fewer weights are concerned compared to the traditional unit selection approaches.

In hybrid systems, a target cost and a join cost are determined for each candidate unit taken from the corpus. The target cost is usually selected from the different combination of: 1) the probabilistic criterion of likelihood of candidate [6]; 2) the Kullback-Leibler divergence (KLD) between target and candidate phone-based HMMs [6]; 3) the difference between the parameter trajectories of candidate and the generated parameter trajectories from HMMS [7]. The join cost uses usually the frame parameterization difference at the point of concatenation or the Mahalanobis distances between the parameterized boundary frames and the corresponding HMM models [6]. Once the cost functions are defined, a lattice of candidates' costs is constructed for each input sentence.

In unit selection stage, Viterbi algorithm (VA) proposed by Viterbi [8] is used to search the optimal path from the lattice. The algorithm can efficiently search for the minimum cost path through a graph. However, the traditional single running Viterbi search has intrinsic problem that all the units selected are of average suitability or quality and some may have larger local errors which can hardly be tolerated.

Various approaches have been proposed in order to solve the problem. In [9], an iterative Viterbi algorithm is proposed and candidates classified as unnatural will be removed between iterations. However, this approach involves an additional prosody model for classification, and iterative search is too time-consuming. Silén, *et al.* [10] introduce the robust Viterbi search to detect the bad units, and replace the bad units using HMM-TTS employing the same parameterization as the unit selection TTS. But the results in [11] show that voice quality is not improved via the mixing of synthetic and real units, and the sudden switching from unit selection to HMM-based synthesis cause unnaturalness which degrades the synthetic speech a lot.

In this paper, we propose to use the hierarchical Viterbi search (HVA) which involves two rounds of Viterbi search. In the first round, a group of sub-paths in the CVS regions will be chosen based on the minimization of local cost function; in the second round, the complete path connecting all the CVS regions will be obtained based on the global minimization of utterance cost. Since the search in each CVS region ignores the other units, the selection of longer continuous speech segments within each CVS region is enabled. Furthermore, the results of first-round search can be used in the second round, which adds no computation complexity to the traditional Viterbi algorithm.

The rest of paper is organized as follows. An overview of our hybrid speech synthesis system will be discussed in Section 2. Section 3 gives the definition and prediction method of CVS region. Afterwards, the use of hierarchical Viterbi

algorithm in unit selection speech synthesis is described in Section 4. In Section 5, several subject experiments are carried out for evaluation. Discussion and conclusion is drawn in Section 6 and 7.

2. Overview of the hybrid system

2.1. Training stage

In this stage, the context-dependent acoustic HMMs are first estimated using acoustic features and label information. Next, a decision tree based model clustering technique is applied to improve the robustness of estimated models. Then the state duration models are also trained. Finally, all the sentences in the speech database are segmented into states using the above acoustic and duration models, and a phone based candidate corpus is constructed.

2.2. Synthesis stage

In the synthesis stage, the input text is first converted into a serial of labels with context features via text analysis. Next, the context dependent acoustic model and duration model of the labels are determined by the decision trees and clustered HMMs. Then the CVS region is predicted using the above directing models.

After that, a lattice of candidates' costs is constructed: the target cost function is taken from [6], while the concatenation cost use directly the frame parameterization difference at the point of concatenation. Our preliminary test indicates that this join cost can enable the selection of longer continuous speech segments for synthesis than the function given in [6].

Finally the proposed hierarchical Viterbi algorithm is applied to select the best group of candidates using the CVS region predicted.

The schematic diagram of our hybrid system is shown in Figure.1.

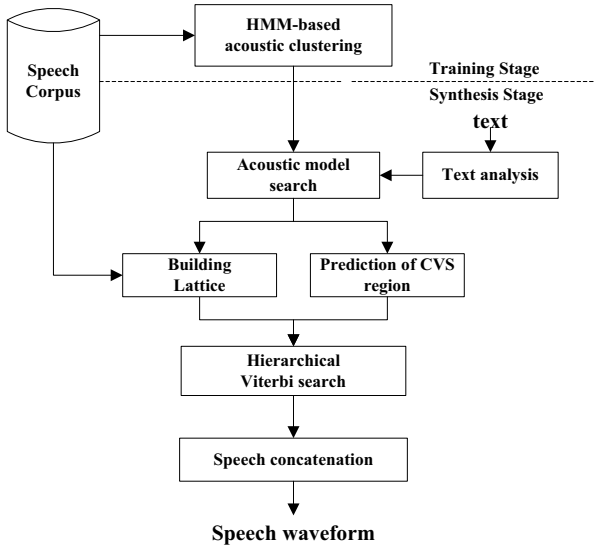

Figure 1: *Schematic diagram of proposed hybrid System*

3. Definition and prediction of CVS

3.1. Definition of CVS

Fig 2 gives an example of pitch contour for the sentence "小学门口的"。The consecutive voiced phones(/ue/, /m/, /en/) are pronounced together without silence, and their pitch contours are continuous in this local region. So these three phones formed a CVS region. In this paper, we defined CVS region as a region which is formed by two or more voiced phones, and whose observation of pitch has a continuous value.

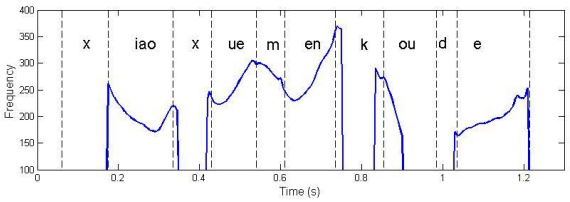

Figure 2: *Pitch contour for "小学门口的"*

In Mandarin, most of syllables are formed by an initial and a final (both referred as a phone in this paper). Some initials are unvoiced like /b/, /p/, /q/; the others are voiced like /m/, /n/, /l/. The finals are all voiced. CVS regions are widely spread in the corpus. Figure 3 gives the ratios of sentences with CVS region in different size of corpus.

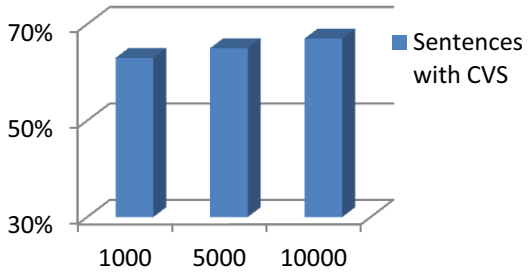

Figure 3: *Ratios of sentences with CVS region in different size of corpus*

In order to explore the significance of the units of CVS region towards the quality of synthetic speech, a pilot experiment is held. 30 sentences, randomly chosen from the holdout corpus, are synthesized using our hybrid system via traditional Viterbi algorithm, and two native speakers are then asked to label the unnatural units, as well as the units from the CVS region. After the labeling, 6 sentences without CVS region are omitted. Table 1 gives the results of the naturalness test results.

Table 1. *Results of naturalness test*

listener	CVS	Total	Ratio
A	37	48	77%
B	25	33	86%

From Table 1, it can be seen that, although the two listeners have different perception of naturalness, they both believe that more than three-quarter perceived unnatural units are from the CVS region. According to the listeners, a tiny

mismatch in the concatenation point within the CVS region can deteriorate the synthetic speech a lot.

3.2. Prediction of CVS

In Mandarin, the voiced/unvoiced condition of certain phone can be approximately predicted by the phonetic rules, so is the location of CVS region. In other languages, however, the rules may not be well summed. For general purpose, we adopt the pitch model to predict the location of CVS region.

While the observation of pitch has a continuous value in the voiced region, there exists no value for the unvoiced region. In the model training stage in 2.1, the MSD-HMM [12] is introduced to model the pitch. After the state decision tree based model clustering, the weights for each space of the clustered model are also trained.

In the unit selection stage, a group of directing models is available. For each model, a voiced/unvoiced threshold value is set to indicate the state is voiced or not. In this paper, if the weight of voiced space is larger than 0.5, then the state is deemed to be voiced. After the voiced/unvoiced condition of all states is checked, another threshold value is implemented to further predict the condition of phone model. If more than half of the states of the model are voiced, then the model is considered as voiced, vice versa. Since 7-state left-to-right with no skip HMM structure is adopted for each phone, this threshold is set to be 3.

Finally, all the models of the group are classified, and two or more consecutive voiced models predict a CVS region. To check the precision, 30 sentences are synthesized and CVS region is predicted, one skillful listener is employed to check the precision from both the sound and the spectrogram. The results are presented in Table 2.

Table 2. *Results of CVS Prediction*

Predicted	Total	Correct Rate
94	100	94%

From the table, it can be seen that the correct rate for the prediction is very high (above 90%), and can achieve precision needed for the HVA proposed in the next Section.

4. Hierarchical Viterbi algorithm for unit selection TTS

4.1. Hierarchical Viterbi algorithm

Assuming the cost function lattice has N phones. For phone i ($i = 1,...,N$), one candidate is u_i . Let $C^t(u_i)$ be a target cost of u_i defined in [5], and let $C^c(u_i, u_{i+1})$ be the concatenation cost between u_i and u_{i+1}, which is given as

$$C^c(u_{i-1}, u_i) = C^c_{lf0}(u_{i-1}, u_i) + C^c_{spec}(u_{i-1}, u_i) \quad (1)$$

, the superscripts of $lf0$ and $spec$ stand for the F0 part and spectrum part. If the utterance has M CVS regions ($r_1,...,r_M \subseteq R$), the Hierarchical Viterbi algorithm can be described as follows.

In the first round, a sub-path is obtained by minimizing the local cost function $C_{cvs,m}$ in each CVS region. For region r_m , $C_{cvs,m}$ is given as

$$C_{cvs,m} = \sum_{i \in r_m} C^t(u_i) + \sum_{i \in r_m \cap i-1 \in r_m} C^c(u_{i-1}, u_i) \quad . \quad (2)$$

In the second round, a global path connecting all the CVS regions is obtained by minimizing the global cost function C_g ,

$$C_g = \sum_{i \notin R} C^t(u_i) + \sum_{i \notin R \cup i-1 \notin R} C^c(u_{i-1}, u_i) \quad . \quad (3)$$

The minimization problem can be solved by Viterbi Algorithm by ignoring target and join costs of the units within all of the CVS Region.

Figure 3 gives an example of HVA. For a CVS region r ($r = \{i-1, i, i+1\}$), the local optimum local unit sequence are searched first, which is shown in Round 1 using gray marks, then the final utterance path is obtained in Round 2 .

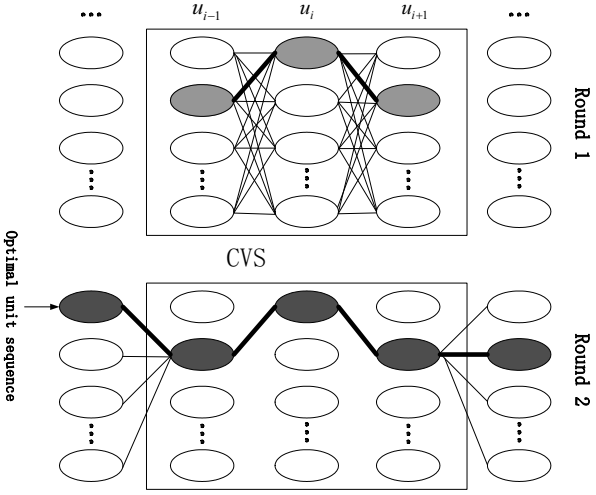

Figure 4: *Hierarchical Viterbi search*

4.2. Complexity evaluation of HVA

The use of HVA does not increase the computation load of the unit sequence selection. The traditional VA algorithm is replaced by a stepwise VA algorithm, and the paths obtained in first-round search are independent with the global path. Furthermore, if there is no CVS Region predicted in the utterance, the HVA will reduce to the traditional VA.

5. Experiments

5.1. Experimental setup

The evaluation experiments are carried out on the hybrid system described in Section 2. The database used for HMM training and unit selection consists of 10000 phonetically balanced Mandarin sentences. Speech signal is analyzed at 5 ms frame shift and LSP order is 24 plus one extra order for energy. 7-state left-to-right with no skip HMM structure is adopted for each initial/final in Mandarin. The acoustic models are trained using HTS 2.2[12]. The same settings of target cost

as [6] are implemented, and the join cost is calculated by equation (1).

In synthesis stage, for each phone, the top 50 candidates are selected for the hierarchical Viterbi search.

For comparison purpose, the traditional VA based system is also constructed, and the settings of cost functions of this system are the same as the above system.

5.2. Subjective Comparison

Synthesis quality of the described hybrid synthesis system was evaluated by two pair wise comparison tests, which includes the "speech quality" and "naturalness". The former measures the segmental articulation and prosodic fluency of the whole sentence and the latter evaluates discontinuity caused by concatenation.

10 randomly chosen sentences with CVS Region were synthesized using HMM-based unit selection (a) in a conventional form using Viterbi algorithm (VA) and (b) in Hierarchical Viterbi algorithm (HVA) as proposed here. 10 native listeners with good skills in Mandarin attended the test.

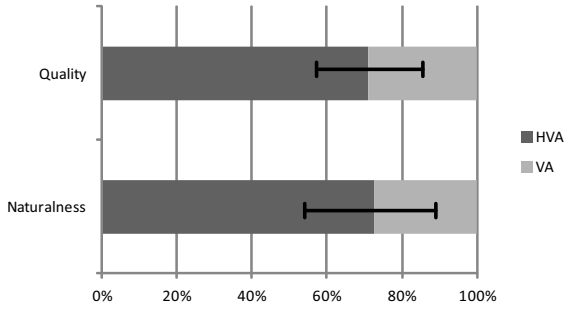

Figure 5: *Preference Results for quality and naturalness*

The results of the comparison tests are presented in Figure 5. It can be seen that the proposed approach was preferred over the conventional VA approach in both the naturalness and the speech quality. For naturalness, the preference percentage of HVA was 72.50% ± 16.38%; for speech quality, the preference percentage of HVA was 71.75% ± 12.45%.

By ignoring the phones out of the CVS region in the first-round search, the most appropriate candidates can be selected within the CVS region where most unnaturalness takes place in. The use of hierarchical Viterbi algorithm minimizes the local cost while increase the global cost slightly compare to the traditional VA. However, the comparison results show that the sacrifice of global cost is worthy since the output speech is more preferable for human perception.

The results also suggest that the human perceptions are especially affected by any local inconsistency (especially in the CVS region), and they score intrinsically different than the cost function does. The algorithm we proposed reduces the local inconsistency and compensates for the lack of cost function.

6. Discussion

Although the output speech has been improved a lot via implementation of the hierarchical Viterbi algorithm compare to the baseline system, there are still occasional glitches or artifacts found within the CVS region. Two reasons emerge from the further examinations.

1) The cost functions need to be better defined. The unnaturalness in synthetic speech may be an inappropriate pitch contour, duration, or power. According to the listeners, most of intolerable unnaturalness is caused by the inappropriate pitch contour in CVS region; however, in the rest part, the inappropriate spectrum or energy are the main cause to unnaturalness. This perceptual difference can be reflected by using different cost functions in each round of Viterbi search. Since the two rounds of Viterbi search are completely independent, different combinations of cost functions can be tried and testified to better reflect the human perception of unnaturalness.

2) The models for the short voiced phones (/l/, /m/, /n/, etc) need to be refined. One obvious problem for the tied model trained by HTS is over-smoothing. For short voiced phones whose context change dramatically from left to right, the perceptual preferable candidates always have a large target cost which exclude themselves during the pre-selection. This problem cannot be rectified by the adjustment of search algorithm since there are sometimes no proper candidates left. The only solution is to overcome the over-smoothing problem and refine the mentioned models.

All of the problems mentioned above will be tasks of our future work.

7. Conclusions

In this paper, an overview of our Mandarin hybrid system is shown first. Next, the definition of the CVS region is given and the importance of naturalness of CVS region towards the quality of whole utterance is analyzed. Then an effective and reliable prediction method for CVS region is proposed and testified. After that, a hierarchical Viterbi algorithm which includes two rounds of Viterbi searching is proposed to replace the traditional Viterbi algorithm. The Evaluations are conducted on the Mandarin corpus, and the results suggest that both the naturalness and quality of the synthetic speech can be improved by including the prediction of CVS region and the implementation of hierarchical Viterbi algorithm which give a priority to the CVS region.

8. Acknowledgements

This work is supported by the Major Program for the National Social Science Fund of China (13&ZD189)", the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, No.61203258, No.61305003, No. 61332017, and No.61375027).

9. References

[1] N. Iwahashi, N. Kaiki, and Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences,* vol. 76, pp. 1942-1948, 1993.

[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings.,*

1996 IEEE International Conference on, 1996, pp. 373-376.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication,* vol. 51, pp. 1039-1064, 2009.

[4] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang*, et al.*, "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007," in *Blizzard Challenge Workshop*, 2007.

[5] R. Zhang, J. Tao, Y. Li, and Z. Wen, "A novel unit selection method for concatenation speech system using similarity measure," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, 2013, pp. 1-5.

[6] Z.-H. Ling and R.-H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. IV-1245-IV-1248.

[7] Y. Qian, Z.-j. Yan, Y.-j. Wu, F. K. Soong, G. Zhang, and L. Wang, "An HMM trajectory tiling (HTT) approach to high quality TTS Microsoft entry to Blizzard Challenge 2010," 2010.

[8] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on,* vol. 13, pp. 260-269, 1967.

[9] D. Lin, Y. Zhao, F. K. Soong, M. Chu, and J. Zhao, "Iterative unit selection with unnatural prosody detection," in *INTERSPEECH*, 2007, pp. 2909-2912.

[10] H. Silén, E. Helander, J. Nurminen, K. Koppinen, and M. Gabbouj, "Using robust viterbi algorithm and HMM-modeling in unit selection TTS to replace units of poor quality," in *INTERSPEECH*, 2010, pp. 166-169.

[11] M. P. Aylett12 and C. J. Pidcock, "The CereProc Blizzard Entry 2009: Some dumb algorithms that don't work," 2009.

[12] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 229-232.