A SMALL FOOTPRINT HYBRID STATISTICAL/UNIT SELECTION TEXT-TO-SPEECH SYNTHESIS SYSTEM FOR AGGLUTINATIVE LANGUAGES

Ekrem Guner, Cenk Demiroglu

Ozyegin University, Istanbul, Turkey

ekrem.guner@ozyegin.edu.tr, cenk.demiroglu@ozyegin.edu.tr

ABSTRACT

Despite its success, unit selection based text-to-speech synthesis (TTS) has has some disadvantages such as sudden discontinuities in speech that distract the listeners. The HMM-based TTS (HTS) approach has been increasingly getting more attention from the TTS research community. One of the advantage is the lack of spurious errors that are observed in the unit selection scheme. Another advantage of the HTS system is the small memory footprint requirement which makes it attractive for embedded devices. Here, we propose a novel hybrid statistical unit selection TTS system for agglutinative languages that aims at improving the quality of the baseline HTS system while keeping the memory footprint small. The intelligibility and quality scores of the baseline system are comparable to the MOS scores of English reported in the Blizzard Challenge tests. Listeners preferred the hybrid system over the baseline system in the A/B preference tests.

Index Terms— speech synthesis, HMM-based TTS, Turkish TTS, small memory footprint, agglutinative languages

1. INTRODUCTION

The HMM-based text-to-speech (HTS) approach has been increasingly getting more attention from the TTS research community. Despite its lower quality compared to the unit selection approach, it has some advantages which make the HTS approach attractive both for speech researchers and speech industry. One of the advantages is the lack of spurious errors that are observed in the unit selection scheme. In fact, in Blizzard Challenges 2005 and 2006, mean opinion scores (MOS) of an HTS system was higher than a unit selection system because of the sudden annoying artifacts in speech generated with the unit selection system [1]. However, in general, unit selection systems can generate higher quality speech than the HTS systems when spurious errors are not present.

A second advantage of the HTS approach is the small memory footprint of the voice database. As opposed to the large databases needed for the unit selection systems, a couple of megabytes is enough to store the voice database for HTS. That advantage is important in some of the embedded applications such as in-car speech interfaces where small memory footprint is often a requirement.

Besides HTS and unit selection approaches, there are hybrid systems that attempt to generate speech that is smooth as in the HTS approach but natural-sounding as in the unit selection approach. Those systems can be divided into several categories. In one approach, parameter training for HTS can be improved by minimizing the error of selecting the wrong unit from the database when the HTS parameters are used to calculate the target costs in unit selection [2],[3],[4],[5]. In another approach, HTS-generated waveforms are interweaved with the speech units selected from the database [6],[7]. The idea is to use smooth HTS-generated waveforms when a unit with a low transition cost cannot be found in the database. In a third approach, synthetic speech generated with unit selection is smoothed at the unit boundaries using an HTS approach [8].

As opposed to most of the existing hybrid methods that are focused on improving the quality of a unit selection system, here, we propose a hybrid HTS/unit selection algorithm to boost the performance of our Turkish HTS system. In the existing hybrid systems, small memory footprint advantage of the HTS system is lost since both a unit selection and an HTS system are used. A key novelty in this work is a hybrid system that keeps the voice database size small while improving the quality of the HTS system. Turkish is an agglutinative language and many different words can be generated from the same root word by using a limited set of suffixes. Given a typical Turkish utterance, a significant number of the words contain one or more suffixes. Moreover, ignoring silences, around one third of the speech is composed of suffixes. In the proposed system, a database for the most frequently occurring suffixes is created in training. In synthesis, best fitting suffixes are selected using the proposed suffix selection algorithms. Then, the selected suffixes are used in HTS within the proposed parameter generation algorithms. Although, the idea is applied to Turkish TTS here, it can also be used for other morphologically rich languages such as Finnish, Estonian and Czech.

This paper is organized as follows. An overview of the proposed hybrid system is given in Section 2. The suffix selection algorithms are presented in Section 3, and the parameter generation algorithms are presented in Section 4. Finally, experiment results of the hybrid system are reported and discussed in Section 5.

2. OVERVIEW OF THE HYBRID SYSTEM

An overview of the training and synthesis phases of the proposed system is shown in Fig. 1. A brief description of the proposed algorithm is given here. Details of the suffix selection and the HMM parameter generation algorithms are given below.

In the training phase, HMM models and a decision tree are generated for the target speaker using speaker dependent training with the HTS tools [9]. Then, a morphological analyzer is used to analyze the words in the speech database. To create a suffix database, waveforms that correspond to the suffixes labelled by the morphological analyzer should be extracted from speech. Forced alignment is used with the speaker-dependent HMM models to align text and speech data. Suffix units are then extracted from the speech signal using the alignment information.

This project is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under Project 109E281

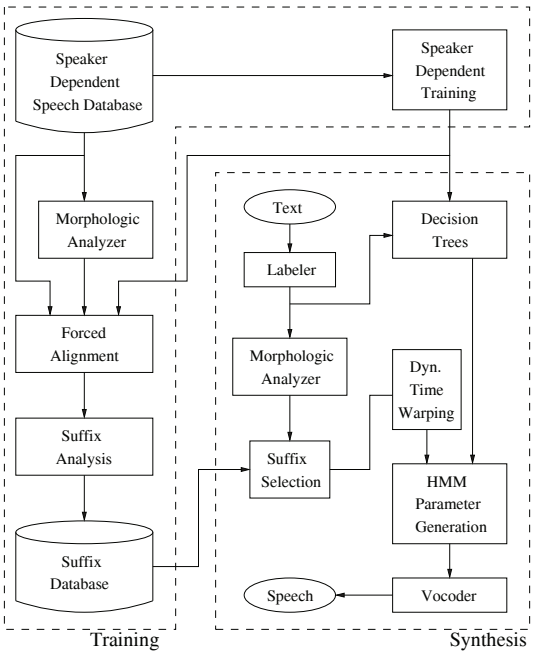

Fig. 1. Overview of the proposed hybrid system

Suffix units are parametrized using LPC analysis and only the LSF and pitch parameters are stored. Besides those parameters, each entry in the suffix database contains a flag that indicates the presence of silence at the right context of the suffix (phrase ending) and another flag that indicates the presence of stress on the suffix. Moreover, beginning and end times of the state-level segments are also stored in the database.

In the synthesis phase, HMM models that correspond to the input text is determined using the decision tree. Input text is analyzed using the morphological analyzer, and, for each suffix in the text, the best fitting suffix is selected using the algorithms described below. The statistics predicted by the decision tree and the parameters of the unit that is selected from the suffix database are combined together and fed into the parameter generation algorithms described in Section 4. Finally, the parameter sequences generated are used in an LSF vocoder to synthesize speech.

The morphological analyzer described in [10] is used here. The analyzer generates the root word and the morphemes of a given word. Both the inflectional and derivational features of the morphemes are produced. Nominal features (case, person/number agreement, possessive agreement) and verbal features (tense, aspect, modality, and voice) are indicated with special tags.

The analyzer sometimes returns multiple alternatives. A morphological disambiguation tool can be used to resolve such cases [11].

3. SUFFIX SELECTION

When synthesizing utterance i, using the morphological analyzer, we determine the set of suffixes $\{S_i^j\}$ in the utterance where $j = 1, 2, ..., N_i$ and N_i is the total number of suffixes in the i^{th} utterance. For the j^{th} suffix, the initial set of available units in the database is denoted by $\{U_1^j\}$. In the initial set, only the stress flag of the suffix is used for selection.

Two different targets are selected for LSF and pitch parameters. Moreover, the cost calculation for those features are also different. The details of the cost calculations are given below.

3.1. Target Cost Calculation

The unit selection algorithm uses a weighted maximum likelihood (w-ML) criterion as the target cost. In the proposed w-ML method, the average log-likelihood for each target is computed by

$$L_{j,k} = \frac{1}{N_{j,k}} \sum_{m=1}^{M} w_{j,k}^m \sum_{f=1}^{f_m} log \left[\frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_m|^{1/2}} \right] - \frac{1}{2}(X_m^f - \mu_m)^T \Sigma_m^{-1} (X_m^f - \mu_m)$$

where $N_{j,k} = \sum_{m=1}^{M} w_{j,k}^m f_m$,

$$w_{j,k}^m = \begin{cases} \gamma_{j,k}^m / f_m & \text{if } m \leq 2 \text{ or } m \geq (M-1) \\ 1 & \text{otherwise} \end{cases} \quad (1)$$

and

$$\gamma_{j,k}^m = \begin{cases} 5 & \text{if } f_m \leq 5 \\ f_m & \text{if } f_m > 5 \end{cases} \quad (2)$$

M is the total number of states, f_m is the total number of frames in state m, Σ_m is the covariance matrix in state m, μ_m is the mean vector in state m and X_m^f is the f^{th} observation of state m. X_m^f contains static, delta, and delta-delta features. The numbers 2 and 5 are found experimentally.

The w-ML measure helps smooth out the concatenation points by assigning higher weight in likelihood computation to states around those points. To further reduce the possibility of a discontinuity artifact, for the pitch parameter, we have used two heuristics to filter out the set of available units in the database for a given suffix. The heuristics are described below.

During parameter generation, the pitch trajectory of the selected unit is time-warped so that it can fit into the synthetic duration estimated with HTS. In our experiments, expanding the pitch trajectory did not cause any audible artifacts. However, compressing the pitch trajectory occasionally caused sudden pitch changes which are perceived as artifacts by the listener. To avoid that problem, the units in $\{U_1^j\}$ that are 20 percent longer than the synthetic duration of the suffix S_i^j are filtered out. The reduced set of units after filtering is denoted by $\{U_2^j\}$.

Pitch fall at the end of phrases is especially important in perception. Thus, if there is a pause or a short pause at the right context of S_i^j, which indicates the end of a phrase, then the units that do not have a pause or a short pause at their right contexts are filtered out from $\{U_2^j\}$, and the reduced set of units are denoted with $\{U_3^j\}$.

Finally, the w-ML criterion is used to select the suffix from $\{U_3^j\}$. The heuristics used in calculating the cost function for pitch are not used for the LSF features. Thus, the w-ML cost is the only criterion in selecting the appropriate suffix in LSFs.

4. PARAMETER GENERATION FOR THE HYBRID SYSTEM

In HTS, each utterance i is composed of a sequence of states and each state s has a set of models $\lambda_{i,s}$ for LSF and pitch features. For LSFs, the probability distribution is defined by a multivariate Gaussian specified by the parameter set $\lambda_{i,s}^{lsf} = \{\mu_{i,s}^{lsf}, \Sigma_{i,s}^{lsf}\}$. For pitch, the probability distribution is defined by a multivariate Gaussian specified by the parameter set $\lambda_{i,s}^{p} = \{\mu_{i,s}^{p}, \Sigma_{i,s}^{p}\}$. These pdf's

are used used to generate the feature trajectories using a maximum likelihood (ML) method.

In the hybrid system, once the best matching suffixes are selected for the LSF and pitch parameters, the features extracted from those suffixes are used to modify the parameter generation process of HTS. Similar to suffix selection, different parameter generation algorithms are used for the two features to obtain the best quality speech. The hybrid parameter generation algorithms for the LSF and pitch features are described below.

4.1. Parameter Generation for LSF

If state s in utterance i occurs within a suffix, the LSF model of the state is updated with $\{\mu_{i,s}^{lsf,suf}, \Sigma_{i,s}^{lsf}\}$ where $\mu_{i,s}^{lsf,suf}$ is the feature vector in the suffix that has the smallest distance to the mean parameter $\mu_{i,s}^{lsf}$. The distance is measured with log-likelihood.

In some cases, the closest frames are not close enough which causes distortion in the synthetic speech. Therefore, a threshold ζ is used to eliminate the distortion in such cases. If the log-likelihood score is below ζ, then $\mu_{i,s}^{lsf,suf} = \mu_{i,s}^{lsf}$.

Even though the mean parameter is selected from real speech, parameter generation process distorts the original spectrum. To avoid that, the variance $\Sigma_{i,s}^{lsf}$ is set to a small number ϵ for the middle frame of state s. Hence, it is ensured that the middle frame is almost exactly the same as the real spectrum of the suffix. This, however, causes discontinuity problems if the state duration is short. To solve the issue, $\Sigma_{i,s}^{lsf}$ is set to ϵ, only if the duration is longer than five frames which is found to be long enough to smooth out the trajectory.

After the pdf's are updated for each state that falls within a suffix, the ML-based HTS parameter generation process is used to create the final LSF trajectories.

4.2. Parameter Generation for Pitch

For the pitch parameters, preserving the intonation pattern of the target unit is important for improving the naturalness of the synthetic speech. Therefore, instead of changing the model parameters of the states, pitch trajectory of the target unit is directly concatenated with the synthetic speech. The pitch trajectory is time-warped to fit into the HTS estimated suffix duration. Every phoneme in the suffix is warped individually as opposed to warping the whole suffix to compensate for the large phoneme duration variabilities in the suffixes.

Directly concatenating the pitch trajectory of the target unit into the HTS generated trajectory creates discontinuities at the boundaries. To address this problem, HTS parameter generation process is used to smooth out the trajectory around the concatenation points. However, besides smoothing the trajectory at the concatenation point, this approach also distorts the target unit trajectory which is undesirable. That problem is solved by setting the variances of the pitch models on the suffix to ϵ. Because the parameter generation process uses an ML-based measure, the system is effectively enforced to preserve the target trajectory while smoothing out the trajectory at the concatenation points.

5. EXPERIMENTS

All systems in the experiments were trained with 30 dimensional vectors consisting of 24 LSFs, 1 log F0 coefficient and 5 voicing strength parameters. Voicing strengths are computed using normalized auto-correlation measure for five evenly spaced spectral bands between 0 and 8000 Hz. Recordings were done in a quiet room with a professional microphone at 44.1 kHz sampling rate. Speech signal is amplitude-normalized and downsampled to 16 kHz before training. Global variance is found to hurt the quality. Hence, it is not used in our tests. Otherwise, default HTS 2.1 toolkit parameters are used in training and synthesis. 70 percent declarative, 15 percent exclamatory and 15 percent interrogative type sentences are used in testing.

500 utterances were recorded by a female speaker. Total duration of the recorded speech is 70 minutes. The female speaker has Istanbul accent.

The baseline system for comparison with the hybrid system uses stress information and CART-tree based pronunciation model. Mixed-excitation is used to reduce the buzziness.

To validate that the baseline HTS system has a state-of-the-art performance, following the Blizzard Challenge approach, Mean Opinion Score (MOS) test is used to test the quality. 10 male and 7 female listeners took the listening tests. All of the listeners were native speakers of Turkish. In the MOS test, subjects were presented 2 sample voices for each MOS score for calibration reasons. Listeners were then presented an utterance and asked to give it a score which represents how natural the sentence sounded. 37 test sentences were selected from news domain and 66 sentences were selected from novel domain. Mean of the MOS scores of the 17 speakers is 3.14 and variance is 1.06. The median of the 17 scores is 3 which is close to the mean.

5.1. Comparison of the Hybrid and Baseline Systems

The same speech database described above for the baseline system is used for the hybrid system.

In order to assess the quality improvement with the hybrid approach, A/B preference test is performed. 50 utterances from a Turkish novel are used in the test. 18 listeners took the test. In 53.9 percent of the utterances, listeners preferred the hybrid system. In 46.1 percent of the utterances, listeners preferred the baseline system. The difference seems significant for Pearson's chi square test at 0.95 confidence level. Thus, there is a preference for the hybrid system over the baseline system. However, the difference is not large. The test results are further analyzed and it was observed that discontinuities that sometimes occur with the hybrid system had an impact on the listener preference. That same effect was also found to have a significant impact on the Blizzard Challenge tests. In fact, some of the HTS systems outperformed the unit selection systems in those tests due to the discontinuity problem [1].

In Turkish, question sentences typically have special suffixes, such as /mi/, /midir/, at the end of the verbs. In some significant number of cases with the baseline system, we have noticed oversmoothed question tags which significantly hurt the listener preference. Most of those issues are resolved with the hybrid system since stress patterns of the question sentences are captured better by the hybrid system. An example case is shown in Fig. 2 where the hybrid system better modeled the pitch rise at the end of a question utterance.

Another interesting syntactic suffix in Turkish is /de/, /da/ which means "also" in English. They are written as if they are independent words while they are treated as a suffix of the word that they come after in this work. Those tags are very commonly used in Turkish and using correct prosody for them is important to convey the correct semantic message. The hybrid system generated more natural prosody for those suffixes since their intonation patterns are selected from the natural units in the suffix database.

The hybrid system improved the intonation contours and the

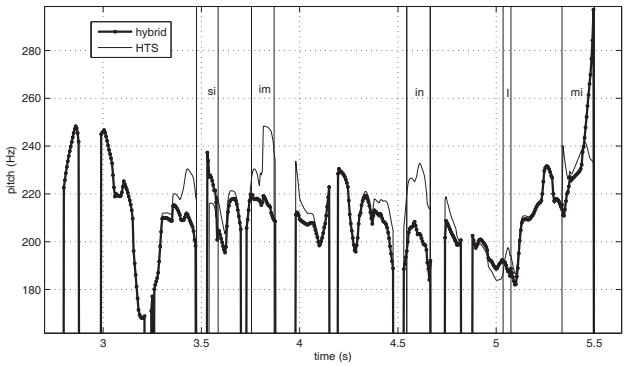

Fig. 2. Comparison of pitch trajectories for the HTS and hybrid systems. Borders of the seven suffixes occurring in the utterances are shown. The final suffix /mi/ indicates a question. Sudden pitch rise that is expected at the end of the question utterance is better modelled with the hybrid system.

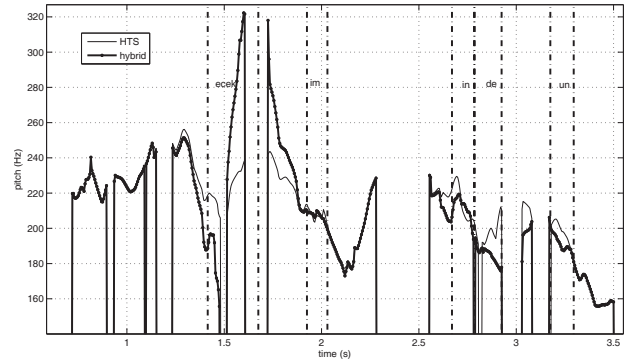

Fig. 3. Comparison of pitch contours for the HTS and hybrid systems. Borders of the five suffixes occurring in the utterances are shown. Sudden pitch variation on the suffix is modeled better with the hybrid system. Synthetic speech with the hybrid pitch contour was perceived as more natural by the listeners.

clarity of suffixes. Analyzing the listener feedback, we have found that improvement with the intonation contours made the most difference in the improved perceptual quality. Another example to pitch contour improvement with the hybrid system is shown in Fig. 3.

The LSF features improved the clarity and reduced the roboticness of the long duration sounds such as long vowels. For the shorter sounds, the effect is less noticeable since the trajectory is smoothed by the parameter generation algorithm.

As opposed to the existing hybrid systems that require substantial amount of storage space for the voice database, the proposed system uses 10MB of memory to store the unit selection database and around 2MB of memory to store the HTS database. Moreover, further significant reduction in the suffix database is possible with vector quantization and other compression techniques. Thus, with the proposed hybrid system, the small memory footprint advantage of the HTS system is maintained while improving the quality.

6. CONCLUSIONS AND FUTURE WORK

In this work, we proposed a novel hybrid HTS/unit selection TTS system for agglutinative languages. The goal of the proposed system is to improve the quality of the baseline system using a suffix selection scheme. As opposed to other hybrid systems literature, the proposed system does not substantially increase the memory requirements for storing the voice database. Therefore, it is suitable for embedded devices with low memory resources. Although the idea is applied to Turkish here, the proposed method can be used for other agglutinative languages also.

In the preference tests, listeners preferred the hybrid system over a state-of-the-art baseline HTS system. We have found that discontinuities at the concatenation points was the main reason when listeners preferred the baseline system. Otherwise, the proposed system generated more clear and natural speech that is less robotic compared to the baseline system. In the future work, we will work on developing smoothing algorithms at the concatenation points to further improve the quality.

7. REFERENCES

[1] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. ICASSP*, 2007, vol. 4, pp. 1229–1232.

[2] H. Lu, Z.H. Ling, M. Lei, C.C. Wang, H.H. Zhao, L.H. Chen, Y. Hu, L.R. Dai, and R.H. Wang, "The USTC System for Blizzard Challenge 2009," in *Blizzard Challenge Workshop*, 2009.

[3] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new TTS from ATR based on corpus-based technologies," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[4] S. Rouibia and O. Rosec, "Unit selection for speech synthesis based on a new acoustic target cost," in *INTERSPEECH*, 2005, pp. 2565–2568.

[5] Y. Qian, Z.J. Yan, Y. Wu, F.K. Soong, X. Zhuang, and S. Kong, "An HMM Trajectory Tiling (HTT) Approach to High Quality TTS," in *INTERSPEECH*, 2010, pp. 422–425.

[6] S. Tiomkin, D. Malah, S. Shechtman, and Z. Kons, "A hybrid text-to-speech system that combines concatenative and statistical synthesis units," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. pp, no. 99, 2010.

[7] V. Pollet and A. Breen, "Synthesis by Generation and Concatenation of Multiform Segments ," in *INTERSPEECH*, 2008, pp. 1825–1828.

[8] M. Plumpe, A. Acero, H.W. Hon, and X. Huang, "HMM-based smoothing for concatenative speech synthesis," in *Fifth International Conference on Spoken Language Processing*, 1998.

[9] Junichi Yamagishi, Heiga Zen, Yi-Jian Wu, Tomoki Toda, and Keiichi Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia, Sept. 2008.

[10] K. Oflazer and S. Inkelas, "A finite state pronunciation lexicon for Turkish," in *Proceedings of the EACL Workshop on Finite State Methods in NLP, Budapest, Hungary*, 2003, vol. 82, pp. 900–918.

[11] D. Yuret and F. Ture, "Learning morphological disambiguation rules for Turkish," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 328–334.