ANALYSIS OF SPEAKER SIMILARITY IN THE STATISTICAL SPEECH SYNTHESIS SYSTEMS USING A HYBRID APPROACH

Ekrem Guner, Amir Mohammadi, and Cenk Demiroglu

Ozyegin University, Istanbul, Turkey

{ekrem.guner,amir.mohammadi}@ozu.edu.tr,cenk.demiroglu@ozyegin.edu.tr

ABSTRACT

Statistical speech synthesis (SSS) approach has become one of the most popular and successful methods in the speech synthesis field. Smooth speech transitions, without the spurious errors that are observed in unit selection systems, can be generated with the SSS approach. However, a well-known issue with SSS is the lack of voice similarity to the target speaker. The issue arises both in speaker-dependent models and models that are adapted from average voices. Moreover, in speaker adaptation, similarity to the target speaker does not increase significantly after around one minute of adaptation data which potentially indicates inherent bottleneck(s) in the system. Here, we propose using the hybrid speech synthesis approach to understand the key factors behind the speaker similarity problem. To that end, we try to answer the following question: which segments and parameters of speech, if generated/synthesized better, would have a substantial improvement on speaker similarity? In this work, our hybrid methods are described and listening test results are presented and discussed.

Index Terms: speech synthesis, statistical speech synthesis, speaker similarity, speaker adaptation, hybrid synthesis

1. INTRODUCTION

The statistical speech synthesis (SSS) approach has become one of the most popular and successful methods in the speech synthesis field. Despite its lower average quality compared to the unit selection approach, it has some advantages which make the SSS 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 the unit selection system because of the relatively lower number of sudden annoying artifacts in speech generated with SSS [1].

In addition to generating smooth synthetic speech, SSS systems have other important advantages. One of the most important advantages is the ability to adapt to a new speaker's voice with a couple of minutes of data. Thousands of voices have been generated with SSS using speech databases prepared for speaker-independent speech recognition systems [2]. However, a problem with the SSS method is the low similarity of synthetic voice to the original speaker. Both speaker-dependent and speaker-adaptation methods produce voices that have low similarity to the target speaker. Moreover, the problem is not significantly alleviated by using more adaptation data from the target speaker [3] which indicate potential inherent bottleneck(s) in the system.

There are many methods that were shown to improve the naturalness of SSS (for example, the global variance (GV) method [4] or the minimum generation error training [5]); however, same progress could not be achieved for the similarity issue. Although the problem is well-known, there has not been any method that could substantially reduce the issue. In one of the recent works, similarity scores have been improved by using higher sampling rates and increasing the pitch variations for a speaker-dependent voice [6]. However, more work is needed both for understanding and solving the issue.

The goal of this paper is to gain more insight to the similarity problem. Two potential sources: inaccurate acoustic modeling and parameter generation algorithms are investigated to understand what are the key factors that distort speaker similarity. We took a novel analysis approach and investigated the problem using hybrid speech synthesis. The core idea of the hybrid methods is to use both unit selection and SSS methods to take advantage of their strengths and get better quality speech than them [7]. Here, we used the hybrid approach to use original recordings in selected segments of speech and generate the rest of the speech with SSS to understand which speech segments and features have the most effect in similarity degradation. For example, in one approach, we focused on transitional segments of speech which are not modeled well with SSS, and used original recordings for those while synthesizing the rest of the speech using SSS. Using five such hybrid techniques, we investigated key parameters that need to be improved in SSS to increase speaker similarity.

This paper is organized as follows. Parameter extraction and modeling is described in Section 2.1. Maximum-likelihood parameter generation algorithm and the hybrid algorithm are described in Section 2.2 and Section 2.3 respectively. The hybrid methods that are proposed to investigate the effects of different factors in degraded similarity are explained in Section 3. Experiment setup and results are presented and discussed in Section 4.

2. OVERVIEW OF THE STATISTICAL SPEECH SYNTHESIS APPROACH

The parameter extraction, acoustic model training, parameter generation and vocoder used in our SSS are described below.

2.1. Parameter Extraction and Modeling

As a first step in training, speech parameters are extracted from the speech database. There are various alternatives for modeling the speech spectrum, such as mel-cepstrum and generalized mel-cepstrum parameters (MGC). Excitation can be modeled with an impulse train for voiced speech and random noise for the unvoiced speech. For the voiced speech, logarithm of the fundamental frequency (LF0) is extracted. One important problem with LF0 is that, although it has continuous values for voiced speech, it is not defined for the unvoiced speech. Therefore, a symbol indicating unvoiced speech is used instead of LF0 for unvoiced speech. That makes the

LF0 a sequence of continous-valued numbers and symbols which is modeled with multi-space distribution (MSD).

The spectrum and pitch features are fused together to create the static feature vector $f_t^T = [c_t^T, (lf0)_t^T]$ at time t. Besides the static features, velocity (Δf_t^T) and acceleration ($\Delta\Delta f_t^T$) features are also used to create the final feature vector $o_t^T = [f_t^T, \Delta f_t^T, \Delta\Delta f_t^T]$. Statistical speech synthesizers can create smooth feature trajectory because of modeling the acceleration and velocity of the features in addition to the static features.

Changes in the pitch contour do not necessarily occur in synchrony with the spectral features. State-level alignments of those two sets of features can be very different. Modeling them jointly results in suboptimal state alignment both for the spectral and pitch features which causes misestimations in modeling those features.To solve that problem, those two stream of features can be trained independently in a multi-stream training framework which provides some flexibility to the system.

Phonemes are modeled with N-state Hidden Semi-Markov Models (HSMM) in the STS approach. As opposed to the HMM approach used in most current speech recognition systems, state durations, $\Pi_i(d)$ are modeled by a Gaussian distribution, $N(\mu_{d,i}, \Sigma_{d,i})$, in the HSMM approach. This allows the flexibity to set and change the phoneme durations explicitly. Spectral features are typically modeled with a multivariate Gaussian distribution $N(\mu_{c,i}, \Sigma_{c,i})$ and pitch features are modeled with multivariate Gaussian distribution $N(\mu_{p,i}, \Sigma_{p,i})$ for voiced states. Acoustic model parameters λ are trained with a maximum likelihood approach

$$\widehat{\lambda} = \underset{\lambda}{\operatorname{argmax}}\, p(O|\lambda). \tag{1}$$

Because there is no closed-form solution, expectation-maximization algorithm is used for estimating λ.

In text-to-speech systems, it is critical to take into account the context of phonemes to synthesize them as close to natural as possible. In the HTS approach, phonemes have labels that contain information about the phone-level, syllable-level, word-level, phrase-level and utterance-level context in addition to syntactic features such as part-of-speech tags and intonation tags such as TOBI endtones for phrases. Although those labels help accurately model the phoneme parameters, it is impractical to collect enough training for each possible combination of different contexts. Therefore, decision trees are used to cluster phoneme states that have different labels but that are automatically found to be similar.

2.2. Parameter Generation and Synthesis

Once the acoustic models are trained, they can be used to synthesize speech for a given text. The first phase of synthesis is to generate a sequence of phonemes from the text with an associated label for each phoneme. Each phoneme is then modelled with an HSMM and the HSMM models are concatenated to represent the final utterance. Because pitch and MGC parameters are modelled independently, separate sequences of HSMMs are used for them. State-id's of the HSMMs are found using the decision trees and phoneme labels described above. Duration distributions for each state is also found using the decision trees trained for the duration parameter.

Once the state-id's, therefore the emission pdf's, for each state is known, the parameter sequence O for MGC or pitch can be generated using

$$\widehat{O} = \underset{O}{\operatorname{argmax}}\, p(O|\lambda) = \sum_{allQ} p(O|Q,\lambda)p(Q|\lambda). \tag{2}$$

where Q represents a possible state sequence for each observation O_t. Eq. 5 can be simplified by choosing the most likely state sequence by

$$\widehat{O} \approx \underset{Q}{\operatorname{argmax}}\, p(O|Q,\lambda)p(Q|\lambda). \tag{3}$$

Eq. 3 can be further simplified by maximizing the state-sequence (state durations) independently. In this case,

$$\widehat{Q} = \underset{Q}{\operatorname{argmax}}\, p(Q|\lambda) \tag{4}$$

and

$$\widehat{O} = \underset{O}{\operatorname{argmax}}\, p(O|\widehat{Q},\lambda). \tag{5}$$

Parameter O contains static, delta, and delta-delta features. However, we are only interested in the static parameters. Therefore, for the spectral features, Eq. 5 can be written as

$$\widehat{c} = \underset{c}{\operatorname{argmax}}\, p(Wc|\widehat{Q},\lambda). \tag{6}$$

where W is used to derive the delta and delta-delta features from the static features. Let $M = [m_{q_1}^T, m_{q_2}^T, ..., m_{q_N}^T]$ and the block diagonal matrix $U^{-1} = diag[U_{q_1}^{-1}, U_{q_2}^{-1}, ..., U_{q_N}^{-1}]$ where $m_{q_i}^T$ is the mean vector and $U_{q_i}^{-1}$ is the inverse covariance matrix of state i. Then, solution to Eq 6 is

$$\widehat{c} = (W^T U^{-1} W)^{-1} W^T U^{-1} M \tag{7}$$

Once the parameters are estimated for MGC and pitch for the whole text, a parametric LPC-based speech vocoder can be used to synthesize the speech. Simple impulse/noise switch typically produces buzzy quality speech. To solve that issue, many systems employ a mixed-excitation approach where impulse and noise are mixed together in different bands. In that case, mixing weights are also estimated and trained in the acoustic model training phase. Similarly, they are generated at the synthesis phase and used by the vocoder.

2.3. Hybrid Approach to Parameter Generation

In the hybrid approach used here, synthesizing speech with natural speech segments injected throughout the utterance is formulated as constrained optimization problem [7]. The objective function in Eq. 6 is maximized with the constraint that $A\widehat{c} = c_{nat}$ where c_{nat} is obtained from a concatenation of static vectors from the natural segments, and A is used to select the natural segments in the speech utterance. Using a Lagrange multiplier approach,

$$\widehat{c} = (W^T U^{-1} W)^{-1} W^T U^{-1} M + (W^T U^{-1} W)^{-1} A^T \gamma \tag{8}$$

where γ is the Lagrange multiplier and can be found by using $\widehat{c}$ in the constraint $A\widehat{c} = c_{nat}$. This method ensures that the trajectory generated follows the natural trajectory when it is available.

3. ANALYSIS OF SPEAKER SIMILARITY USING THE HYBRID APPROACH

Although synthesized speech with SSS is intelligible and smooth, its similarity to the original speaker is low. There are three possible factors that can cause the problem. The first factor is the lossy parametric vocoding technique used in SSS. When speech is synthesized with the correct parameters derived from original recordings, similarity to the original speaker is exceedingly high compared to speech

synthesized with synthetically generated parameters. Therefore, this factor was not tested in our listening tests.

The second potential factor in distortion is the inaccurate acoustic modeling technique used in SSS. It is well-known that HMMs cannot model rapid speech transitions well. Moreover, assuming stationarity on each state is merely an approximation that can cause monotonic speech for long states. Furthermore, in most SSS systems, speech is represented with a single Gaussian which is known to be a crude representation as large vocabulary speech recognition systems typically use at least 32-64 Gaussians per state. Therefore, we investigated the inaccurate speech representation as a potential source of problem in speaker similarity.

The third factor is the parameter generation algorithm. It is well-known that the parameter generation algorithm described in Section 2 generates overly-smooth trajectories. Maximizing the objective function in 6 is equivalent to the least-norm solution because of the Gaussianity assumption which favors smooth parameter trajectories. To solve the issue, global variance (GV) approach has been proposed which adds an additional term to the objective function to reduce the cost of having higher delta and delta-delta parameters. Although global variance increases the overall quality, improvement in speaker similarity with it has not been investigated in detail in the literature.

Here, we focused on the effects of inaccuracies in acoustic modeling and parameter generation algorithms as two potential sources that reduce the speaker similarity. The five different synthesis experiments that are conducted to test those effects are described below.

In all hybrid methods, the original recording with the HMM states first time-aligned using forced alignment. Then, during parameter generation, the system is enforced to use those time alignments. This allowed us to one-to-one match the frames between original and synthesized speech. Moreover, the biasing effect of inaccurate state durations are eliminated from the test and we could focus only on the MGC and LF0 related changes.

3.1. Hybrid Mid-Frame (HMF) Approach

In this approach, we fix the feature vectors that are in the middle of each state to the corresponding vector in the original recording. The goal in this approach is to answer this question: Can the parameter generation algorithm do a good interpolation if the mean vectors are correct on the middle state of each frame. In general, good interpolation performance is obtained but there are problems with the long states. As soon as the state duration increases beyond a certain point, the hybrid MGC and LF0 trajectories approach the SSS generated trajectories as shown in Fig. 1. The effect is less noticeable in the LF0 trajectories.

3.2. Hybrid Long State (HLS) Approach

The HMF approach is not good at accurately modeling the parameter trajectories on the longer states. Smooth trajectories generated with SSS become virtually flat for those long states which occupy a significant portion of speech. For example, stable segments of long vowels typically have long mid-states as shown in Fig 3. Such flat trajectories are expected to increase robotic quality and potentially reduce the speaker similarity. From a different perspective though, those states in general are not expected to model rapid speech transitions. Hence, unnaturally smooth trajectories may not affect the similarity as long as formant locations and bandwidths are modeled correctly on average. Moreover, some of those longer states correspond to fricatives which do not play a significant role in speaker similarity.

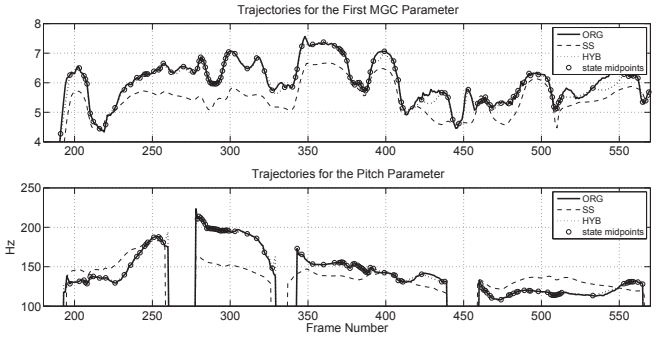

Fig. 1. Comparison of LF0 and the first MGC trajectory for statistical synthesis (SSS), original recording, and the HMF approach. Features of the middle frame of each state is fixed to its corresponding frame in the original recording.

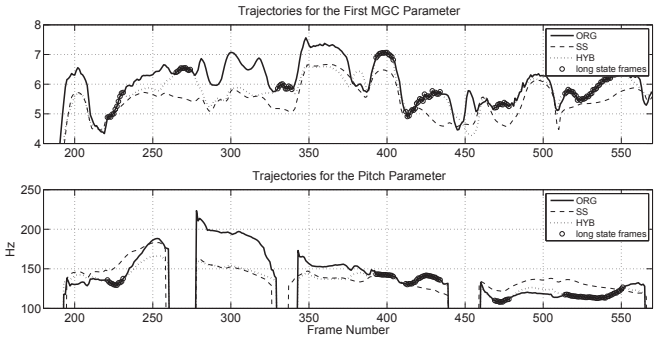

Fig. 2. Comparison of LF0 and the first MGC trajectory for statistical synthesis (SSS), original recording, and the HLS approach. Long states are replaced with original recordings and the rest of the speech is generated with SSS.

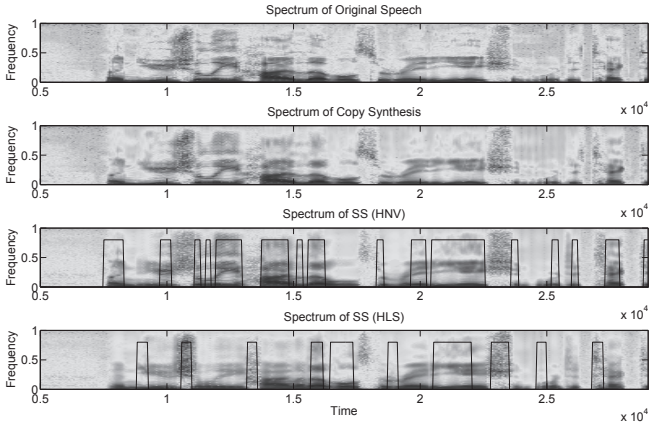

Fig. 3. Comparison of the spectrograms of natural speech, copy-synthesized speech, synthetic speech with HNV, and synthetic speech with HLS

In the HLS approach, the question of how much of the speaker similarity is lost because of unnatural feature trajectories on the long states is investigated by replacing the features on those states with the natural segments using the hybrid approach. In the example shown in Fig 2, the deviation from the SSS trajectory is less with HLS compared to HMF. That is expected since the trajectory is not sampled as frequently as HMF.

3.3. Hybrid Natural Vowel (HNV) Approach

In the HLS approach, relatively stable portions of long speech segments are replaced with original recordings. However, that method does not take into account the beginning and end parts of sounds where a lot of the rapid formant transitions occur. Because vowels are very important in perception and quality, and they typically contain both long stable states and rapid transitions, as shown Fig 3, their parameters are replaced with the natural parameters using the hybrid approach.

The question that is explored with this approach is whether modeling not only the stable mid-vowel segments but also the beginning and end parts of vowels accurately makes a difference in speaker similarity or not. Formants and formant bandwidths are directly related to vocal tract and they have an immense effect on the perception of speech. Modeling them with correct transitional behavior throughout all vowels is expected to increase speaker similarity.

3.4. Hybrid Formant Transitions (HFT) Approach

In the HFT approach, the goal is to replace the transitional segments of speech with the natural speech segments independent of the sound class. The transitional segments in speech is detected by monitoring the formant transitions. The first formant transitions are not always easy to detect and the third formant is not easy track. The second formant, however, typically moves fast in the transitional speech segments and is usually easier to detect. Therefore, the second formant trajectories are detected using the Praat tool, and the rapid second formant changes are detected over a window of N frames. If the median delta feature over N frames centered around the target frame is above a threshold, the frame is labeled to be within a transitional segment. This simple detector was found to perform well in detecting the transitional segments.

3.5. Hybrid Pitch Transitions (HPT) Approach

Rapid pitch transitions are also problematic in the SSS approach. Typically, speech produced by SSS has monotonous pitch trajectory which mostly cannot model the pitch variations in the target speaker's natural speech. In some cases, for example accented speech, modeling the pitch variations correctly can have a significant impact on the speaker similarity. Using natural pitch trajectories where pitch changes rapidly allowed us to explore how much reduction in those variations degrade the speaker similarity.

4. EXPERIMENTS

4.1. Experiment Setup

All systems in the experiments were trained with 75 dimensional vectors consisting of 24 MGCs, 1 log F0 coefficient and their delta and delta-delta parameters. 20 msec analysis window with 5msec frame rate is used for feature extraction. Phonemes are modeled with 5 state HSMMs.

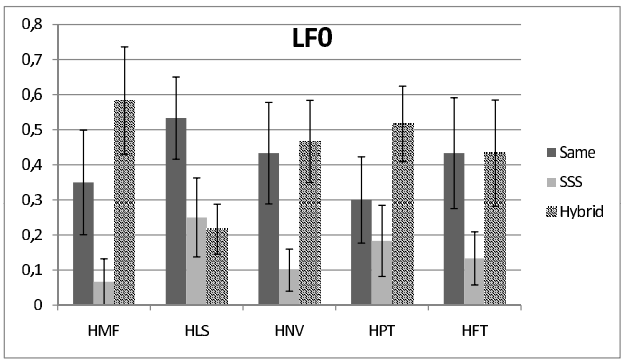

Fig. 4. XAB test results when hybrid LF0 is generated with the five methods.

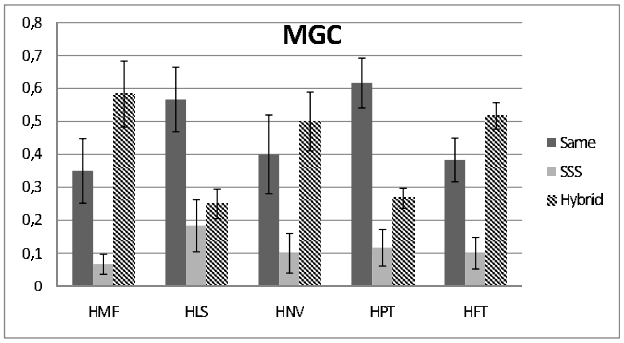

Fig. 5. XAB test results when hybrid MGC is generated with the five methods.

Wall Street Journal (WSJ) database is used to train the average voice and the speaker-adapted voices. Four male speakers with 1250 utterances for each of them are used for training the average voice. Four different speaker-dependent models are generated using CSMAPLR adaptation with an additional MAP step using 1200 adaptation utterances per speaker [8]. HTS 2.2 training and synthesis tools are used to generate the samples for the baseline systems (http://hts.sp.nitech.ac.jp/).

XAB tests are used to measure the similarity of synthesized speech to speaker's voice. Listeners chose if sample A or sample B is more similar to copy-synthesis sample X. Instead of original recordings, their copy-synthesis versions are used where parameters are extracted from the recordings and then speech is resynthesized with those parameters using the HTS vocoder. The purpose of this method is to eliminate the speaker similarity loss that is related to vocoding since the goal in this work is to investigate the effects of acoustic modeling and parameter generation related similarity loss. Six speakers took the test. 36 samples are generated for each hybrid approach with 12 samples for hybrid LF0, 12 for hybrid MGC, and 12 for both LF0 and MGC parameters. Listeners scored a total of 180 samples.

4.2. Results and Discussion

Results of the similarity tests are shown in Figs. 4-6. In Fig 4, the effect of hybrid LF0 approach is shown on similarity preference. Although, hybrid systems seem to outperform the baseline system, variances are very high. Analyzing the results and talking with the

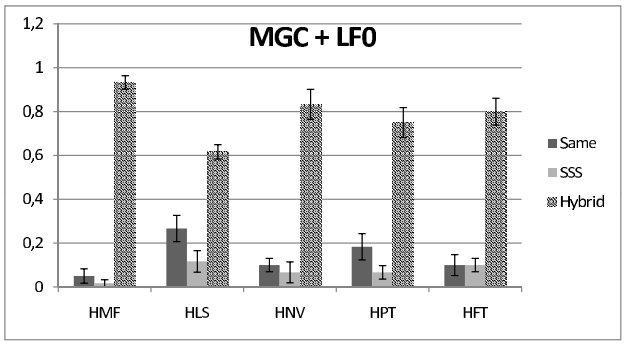

Fig. 6. XAB test results when hybrid MGC and LF0 is generated with the five methods.

listeners, the variance in preference seems to be related to speaking style changes. Some speakers speak with a certain distinctive style which is characterized with pitch variations. When that is the case, listeners preferred the hybrid system which could more successfully mimic those pitch patterns. In some cases, however, reduction in pitch variations and stress simply made the speaker sound more tired, fatigued, and less energetic individual without altering the characteristic properties of his speech. These cases were perceived as just a change in speaker style by some listeners. This has been found to be the major factor in high variability in preferences for the hybrid LF0 case.

Especially for the HPT and HLS cases, listeners put more preference on the hybrid LF0 approach. The HPT approach was expected to perform well since pitch variations are captured accurately with it. Good performance of HLS was found to be related to reduction in pitch monotonicity on the long states.

Results of listening tests with MGC is shown in Fig 5. Hybrid methods also give good performance for MGC especially for the HMF case which was expected because of high sampling rate of the original trajectories. HNV and HFT methods also performed well for this case. Both of those methods involve capturing the rapid transitions in speech as opposed to HLS, which performed poorly, that captures the stable regions. This result indicates that capturing rapid transitions in speech is important for improving the speaker similarity.

Results of the listening tests with hybrid MGC and LF0 are shown in Fig 6. Not only the hybrid approach significantly outperformed the baseline system substantially in all cases, confidence regions are also tighter compared to MGC-only and LF0-only cases. Clearly, similarity to the target speaker improves substantially when both MGC and LF0 parameters are improved at the same time. Performance is best with HMF and worst with HLS. That result was expected given the high sampling rate of HMF and relative insignificance of stable long states in capturing the speech dynamics.

In our analysis, we have observed that MGC related changes improves formant bandwidths. Muffled and robotic quality of speech goes down with reduction in formant bandwidths which not only improves the quality but also the speaker similarity. That fact is also used in improving the quality of speech by postfiltering which is commonly used in SSS systems.

5. CONCLUSIONS AND FUTURE WORK

Hybrid speech synthesis approach was used to analyse the effects of improving different segments of synthetic speech using original recordings. The goal was to invesigate which segments/parameters in speech, if improved, can have substantial impact on similarity. Hybrid LF0 synthesis was sometimes perceived simply as style related changes but not a change in the speaker's voice character. However, in some other cases, LF0 was found to be a distinctive feature and hybrid LF0 was prefered. MGC related changes improve the quality but speakers were specifically asked not to prefer based on quality. However, some of the quality improvements such as reduction in formant bandwidths and more natural formant transitions also heavily affected listeners' preference on similarity. The biggest gain in performance was clearly obtained when both MGC and LF0 segments were generated with a hybrid approach. The improvement was found to be surprisingly consistent over the five different methods. In the future work, we will investigate the substantial boost in similarity when MGC and LF0 are improved together and try to uncover the acoustic-phonetic and perceptual reasons behind it.

6. REFERENCES

[1] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4.

[2] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y.J. Wu, et al., "Thousands of voices for HMM-based speech synthesis-analysis and application of TTS systems built on various ASR corpora," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 984–1004, 2010.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[4] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE transactions on information and systems*, vol. 90, no. 5, pp. 816, 2007.

[5] Y.J. Wu and R.H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006, vol. 1.

[6] J. Yamagishi and S. King, "Simple methods for improving speaker-similarity of HMM-based speech synthesis," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4610–4613.

[7] 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.

[8] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 6683, 2009.