INTERSPEECH 2009 BRIGHTON

Local minimum generation error criterion for hybrid HMM speech synthesis

Xavi Gonzalvo[1,2], *Alexander Gutkin*[3], *Joan Claudi Socoró*[2], *Ignasi Iriondo*[2], *Paul Taylor*[1]

[1]Phonetic Arts Ltd., Cambridge, United Kingdom
[2]Grup de Recerca en Processament , Universitat Ramon Llull, Barcelona, Spain
[3]Yahoo! Europe, London, United Kingdom
xavi.gonzalvo@phonetic-arts.com

10.21437/Interspeech.2009-139

Abstract

This paper presents an HMM-driven hybrid speech synthesis approach in which unit selection concatenative synthesis is used to improve the quality of the statistical system using a Local Minimum Generation Error (LMGE) during the synthesis stage. The idea behind this approach is to combine the robustness due to HMMs with the naturalness of concatenated units. Unlike the conventional hybrid approaches to speech synthesis that use concatenative synthesis as a backbone, the proposed system employs stable regions of natural units to improve the statistically generated parameters. We show that this approach improves the generation of vocal tract parameters, smoothes the bad joints and increases the overall quality.
Index Terms: speech synthesis, HMM, unit selection, hybrid

1. Introduction

Current state-of-the-art text-to-speech (TTS) systems often produce intelligible and natural speech. The most popular approaches are the concatenative and the statistical synthesis [1].

Concatenative speech synthesis is based on the selection and concatenation of recorded units. It is the most widely used approach because it can produce high-quality speech. However, its drawback is that the quality can degrade if, for some reason (e.g., data sparsity), an incorrect joint is produced.

Statistical approach, on the other hand, is based on generating parameters from a trained model. It overcomes the problems that concatenative approach suffers from by producing a smoother output. However, the resulting speech quality is lower than the quality of concatenative systems due to the vocoding. Nevertheless, statistical approach is very attractive because it offers a range of model manipulations such as speaker adaptation.

Hybrid speech synthesis evolved as an alternative to the aforementioned systems. As its name indicates, this approach attempts to combine the benefits of concatenative and statistical synthesis. Several ways [2] to achieve this goal are:

- **Target prediction:** Parameters generated from the HMM can: (a) constraint the features of the target units, such as prosodic parameters [3], (b) select 5 ms segments [4] from acoustic targets or (c) be used in a cost function [5]. It has also been shown [6] that emission probabilities of the HMMs trained using Maximum Likelihood (ML) criterion can be used as target and joint costs in guiding the selection of phone-sized candidate units. In addition, a Minimum Selection Error criterion was proposed in [7] to minimize the number of different units between the selected and natural phone sequences using a generalized probabilistic descent algorithm.
- **Unit Smoothing:** HMMs are used to alleviate the degradation of quality that is due to the use of traditional spectral techniques for smoothing the joints between the units. In one method proposed in [8], the HMMs are used to smooth the spectrum according to what was observed at the junctions of real speech during training while retaining a filter calculated from an actual speech utterance. Another interesting approach reported in [9] fuses the units by imposing dynamic constraints.
- **Unit Mixing:** The approaches in this category mix natural units with the sequences generated by HMMs. This is achieved by concatenation [10] or by using a hybrid voice conversion [11]. In the former case, a multiform segment algorithm determines the optimal sequence of segments (i.e., natural or HMM-generated units) by minimising the degradation of speech. In case the required phoneme context is missing, the conversion method employs combination of unit selection with spectrum generation using speaker-adapted HMMs.

One common characteristic of the hybrid systems described above is that they are driven by concatenative approach, since the latter guarantees the naturalness of the synthetic speech. Moreover, they can help to reduce the effects of data sparsity, decrease corpus footprint and improve the performance of the search algorithms. However, the advantages of the pure statistical approach, such as smooth spectral transition, are lost.

The proposed alternative hybrid HMM-driven approach takes into account the following constraints:

- The synthesis employs HMM parameter generation.
- The synthesis process should avoid mixing units (described above) as this can result in quality degradation that is due to their different spectral nature.
- Concatenative mechanism should be used to improve the quality in stable regions. This is achieved by introducing a weight function $w(f)$ defined over frames f that controls the contribution of concatenation.

An HMM-driven hybrid system sacrifices part of the quality of the natural units, but creates a more flexible structure able to improve the quality of state-of-the-art pure HMM systems while keeping their main advantages. By the definition of these constraints, the resulting system cannot be described as a unit selection smoother because in the limit, the speech would constitute a vocoded version of the concatenative system.

This paper is organized as follows: section 2 introduces the proposed system (training and synthesis stages), section 3 and section 4 describe the LMGE algorithm and the weight function

respectively and section 5 presents the experiments results. Section 6 is the discussion and Section 7 concludes by presenting a summary of our findings and outline of the future work.

2. The Overview of Hybrid System

The proposed hybrid system is shown in Figure 1. It is composed of two modules: one is HMM-specific and another is concatenative. Both components are phone-based.

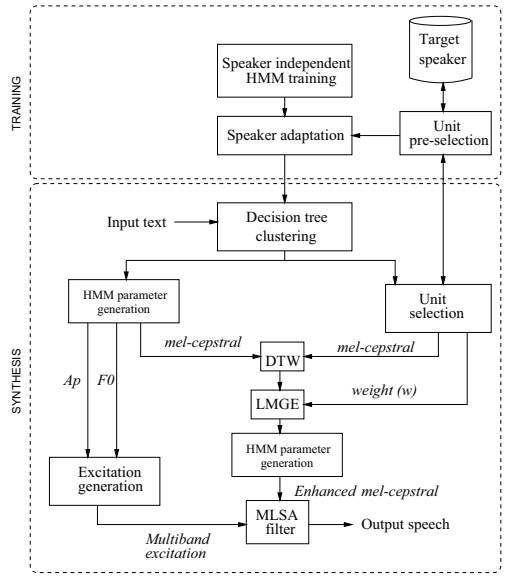

Figure 1: *The hybrid system work flow.*

The training stage consists of two steps. First, an average multi-speaker HMM voice is adapted to target speaker data. The next step builds the concatenative representation. This step essentially constructs an HMM-based unit selection system that uses decision trees produced during the HMM training in the first step. The aim of the second step is to produce the best sequence of natural units that will be used by the HMM to enhance the vocal tract parameters. Since both training steps use the same decision tree based clustering, low spectral distortion is expected.

During the synthesis stage, the text to be synthesized is converted to corresponding linguistic specification. Decision trees are then used to select the corresponding HMM units.

First, durations from the HMMs are used to generate the parameters of vocal tract, F0 and excitation. The natural units are then selected from the decision trees by the concatenative module. A weight function is generated considering the concatenation boundaries. In order to guarantee an optimal synchronization between vocal tract sequences, mel-cepstral information generated by the concatenative and HMM components are aligned using DTW (Dynamic Time Warping) algorithm.

Next, the means and variances of the current HMM units are updated using Local Minimum Generation Error (LMGE) criterion (described in more detail in section 3 below) based on the aligned mel-cepstral sequence of natural units and the weight function. Finally, synthetic speech is generated from mel-cepstral coefficients using MLSA filter driven by multiband excitation. Next sections describe the above steps in detail.

2.1. Speaker-independent HMM training

An average voice model is trained using a set of different speakers in order to obtain a robust voice with only some samples of the target speaker.

At the core of an HMM module we employ a standard mel-cepstral HMM system [12] that uses the latest quality improvements. STRAIGHT vocoding [13] is used for the multiband excitation to reduce the buziness of the vocoder. Global variance is used to alleviate the effects of over-smoothing [14]. We also use explicit state duration probability density function (HSMM) [15].

The model training for HMM-based synthesis constructs a set of context-dependent HMM models, where the vocal tract, pitch and excitation parameters are simultaneously modelled. The high quality speech vocoder (STRAIGHT) is used to analyse the spectral envelope (39-th order mel-cepstral coefficients including delta and delta-delta features) and to get the excitation parameters as an aperiodicity measurement in 5 sub-bands (0-1, 1-2, 2-4, 4-6 and 6-8 kHz). Pitch information (represented in logarithmic domain) is modelled by an MSD (Multi-space probability distribution). Durations use HSMMs to explicitly model state duration probability density functions. In addition, we also train a global variance model. Decision trees are constructed using Minimum Description Length (MDL) criterion. More details can be found in [16].

The adaptation uses Constrained Maximum Likelihood Linear Regression (CMLLR) technique. Mean and covariance matrices are obtained by simultaneously transforming all the parameters. In addition, we employ single bias removal and maximum a posteriori criterion [17]. The adaptation process is attractive because it does not require a huge amount of data and reliable quality can be obtained even with a few adaptation utterances.

2.2. Concatenative system

The aim of the concatenative module is to provide the best natural units to the hybrid system in order to improve the HMM-generated parameters. The method we employ is based on a simplified version of the HMM-based unit selection system described in [6]. The processing consists of two main stages. During the first stage, Kullbak-Leibler Divergence (KLD) measure is used by unit pre-selection algorithm in order to restrict the search space. The leaf nodes of the decision trees obtained during the HMM training (described in previous section) are used to provide target cost estimates. The second stage consists of a classical unit selection search that is carried out to get the best acoustic joints.

3. Local Minimum Generation Error

Minimum Generation Error (MGE) criterion was first introduced in [18]. The idea is to minimize the error of the HMM-generated parameters with respect to the original training data. This is achieved by post-processing of the previously trained models using the standard Maximum Likelihood (ML) criterion and a distance definition.

The problem with the above approach is that when it is applied to a full corpus during the training stage, updating from multiple files smoothes the HMM parameters too much. To alleviate this problem we propose an alternative application of MGE – the Local Minimum Generation Error (LMGE) criterion, which is essentially an MGE that is applied to a single target utterance during the synthesis, rather than the training

stage. By using the LMGE during synthesis time with an optimal single sequence of natural units generated by the concatenative module, it is possible to update the HMM model and adapt its parameters to the current utterance.

First, it is necessary to measure the distortion between the original ($\mathbf{c}$) and the generated parameter ($\tilde{\mathbf{c}}$) vectors. In our case, the original sequence is substituted by the natural units selected from the concatenative module. The similarity metric we adopt is the Euclidean distance:

$$\ell(\mathbf{c}, \lambda) = D_c\left(\mathbf{c}, \tilde{\mathbf{c}}\right) = \left\|\mathbf{c} - \tilde{\mathbf{c}}\right\|^2 \tag{1}$$

The MGE updating rule is given by

$$\lambda_{update} = \lambda_{old} - \varepsilon \sum_{n=1}^{M} \frac{\partial \ell(\mathbf{c}_n; \lambda)}{\partial \lambda}\bigg|_{\lambda=\lambda_{old}} \tag{2}$$

where $\mathbf{c}_n$ is the n training observation, M the number of observations, λ the HMM to be updated and ε the step size for parameter updating. By simultaneously using all the training samples of the sequence produced by concatenation and the updating rule of equation 2, it is possible to deduce the *simplified* equations 3 and 4 to update the mean and the variance [18]:

$$\mu_S = \hat{\mu}_S - \frac{1}{N_{ij}} \sum_{f=1}^{F} (1 - w(f)) \cdot D_{f,k} \tag{3}$$

$$\sigma_S^2 = \hat{\sigma}_S^2 - \frac{1}{N_{i,j} \cdot \hat{\sigma}_S^2} \sum_{f=1}^{F} (1 - w(f)) \cdot D_{f,k} \cdot (\tilde{c}_{f,k} - \hat{\mu}_S) \tag{4}$$

where $D_{f,k} = (\tilde{c}_{f,k} - c_{f,k})$, $N_{i,j}$ is the total number of samples in distribution j in state i, $f \in [1, F]$ is the frame being analyzed, k is the order of the coefficient of the multivariate Gaussian distribution, $\mu_S = \mu_{i,j,k}$ and $\hat{\mu}$ and $\hat{\sigma}$ are the mean and variance prior to update and w is the weight per frame.

Note that the use of LMGE does not convert the HMM generated sequence into natural units because the updating process is performed on the mean and variance of the HMM.

4. Weight Function for Region Updates

In order to select the regions to be updated we introduce a special weight function. In general, the weight function can be used for updating segments, phonemes, words or even full utterances. In this paper, however, we introduce a weight function that operates on a per-frame basis. We define $w(f) \in (0, 1]$ over frames f. The goal of this function is to weight different regions and give higher preference to those regions that do not contain concatenation points.

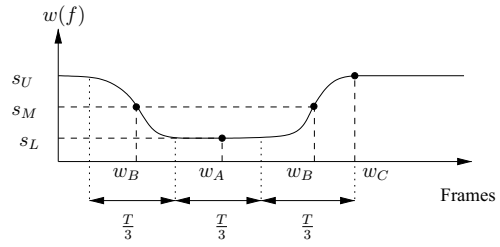

Figure 2: *Per-frame weight function.*

In the upper limit ($w \to 1$), the weight refers to a concatenation point, so HMM does not need to be updated, and in the opposite limit ($w \to 0$), the concatenative data is considered to be stable and as a result, LMGE is applied. Because the weights are defined per-frame, transitions between them are smoothed using a sigmoid representation (see Figure 2). In other words,

$$w(f) = \begin{cases} \frac{a}{b + \mathrm{e}^{-(t+f)}} + s_L & f \neq w_C \\ s_U & f = w_C \end{cases} \tag{5}$$

where,

$$b = \frac{s_M - s_L}{s_U - s_M} \qquad a = b \cdot (s_U - s_L) \tag{6}$$

are defined with regard to the sigmoid limits ($s_U \leq 1$ denotes the upper limit, $s_L > 0$ the lower limit and s_M is the middle point). Figure 2 shows an example region, where T is the duration of a phoneme, w_A is the stable concatenation region, w_B is a sample smoothed transition and w_C is the concatenation point. For each of these locations, the weight function defines the contribution of the concatenative system.

5. Experiments

For our experiments, speaker independent training involves 8 speakers and a total of 21.2 hours of speech. Target adaptation data consists of a 50 minutes recording of a male speaker. This training corpus includes 1000 open-domain utterances automatically segmented. During the pre-processing stage we extracted log F0 information, STRAIGHT mel-cepstral coefficients and five-band aperiodicity components.

The parameters of the weight have been empirically fixed to the following values: $s_U = 0.2$, $s_M = 0.1$ and $s_L = 0.05$. This configuration benefits the use of the concatenative system while smoothing the concatenation points. Note that for smaller amounts of target speaker data s_U may need to be increased.

HMM topology is the standard five-state left-to-right structure with no skips. Each state output probability density functions consists of five streams, similar to the ones used in the Nitech-HTS system [16]. Information about the global variance is also used during synthesis. Hereafter, we refer to three systems: the standard statistical HMM system that is used as a "backbone" of the hybrid system, the concatenative synthesis system that is used to select the natural units to improve the quality of the latter and, finally, the proposed hybrid system that combines previous two approaches.

Figures 3 and 4 show examples of parameters trajectories generated by the three systems. Same phoneme duration was used for the three systems. Figure 3 shows a sequence of the 3rd mel-cepstral coefficient extracted from the three systems. It can be observed that the mel-cepstral sequence generated by the proposed hybrid system is closer to the sequence produced by the concatenative system, except in the region of the concatenation point, where the weight decreases and the sequences overlap with the HMM trajectory.

Sample spectra generated by the three systems are shown in Figure 4. The proposed hybrid system make spectral peaks much sharper than those generated by the HMM, except in the concatenation point where the spectrum is made more similar to the "default" HMM spectrum, resulting in a smoother joint.

We conducted an AB test [1] in order to evaluate the performance of the proposed hybrid system. In this test, 8 listeners were presented with 24 utterances randomly chosen from the test set. The results, shown in Table 1, indicate that the proposed approach increases the average quality of synthesized speech (A denotes the HMM system and B denotes the hybrid system).

[1] Samples can be found at http://www.salle.url.edu/~gonzalvo/hmm/

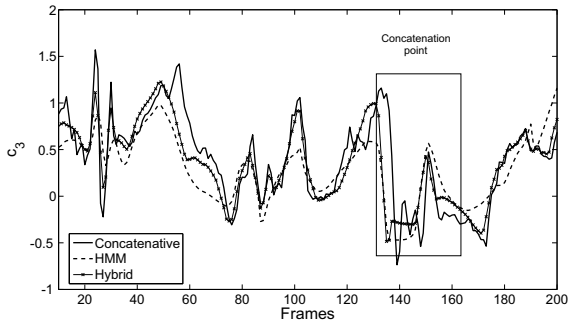

Figure 3: *Mel-cepstrum sequences for the 3rd coefficient.*

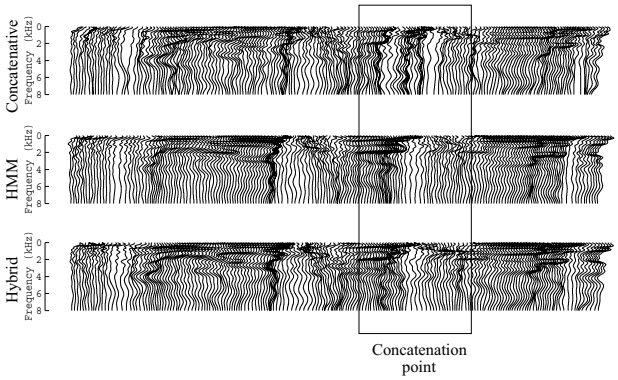

Figure 4: *An example of generated spectrum sequences.*

6. Discussion

Although the hybrid system updates the mel-cepstral parameters, the rest of parameters (F0 and aperiodicity) remain unmodified. In addition, (1) With respect to the F0 updating, we often observed the cases where overall utterance intonation is degraded since concatenative system is not intended to produce natural expressiveness but to select the optimal sequence of natural units. (2) The aperiodicity is modelled in five subbands. However it is actually a continuous representation of the length of the STRAIGHT spectrum which results in undesirable artifacts if it is used within the same updating rule. (3) We anticipated a better quality improvement than the one obtained, hence the simplified assumptions of equations 3 and 4 should be revised in order to obtain a better updating rule.

7. Conclusion

We have presented an HMM-based hybrid system that uses concatenative synthesis in order to improve the quality of the vocoded speech. The system is based on the LMGE algorithm that updates the HMM parameters in order to generate more natural speech. Additional improvement to naturalness is due to a weight function for smoothing the errors between the joints. Moreover, the proposed hybrid algorithm enjoys all the benefits of the standard HMM systems such as robustness and availability of flexible adaptation techniques. In addition, because the HMM system is used as a "backbone", the data sparsity problem is less severe when compared to the concatenative systems. Despite promising results, there are still several issues to be addressed. We need to design a more efficient concatenative synthesis component that will provide more detailed information to the generation module. For example, residual information can be used to improve the excitation or the target cost function can influence the behaviour of the weight function.

Table 1: *AB test for the HMM and hybrid systems.*

A>>B	A>B	A=B	A<B	A<<B
6.8 %	15.98 %	13.88 %	48.86 %	14.48 %

8. References

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

[2] H. Zen, "Statistical parametric speech synthesis," Talk at MIL Speech Seminar, Cambridge University, 2008.

[3] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "XIMERA: A new tts from atr based on corpus-based technologies," in *Proc. of SSW5*, 2004, pp. 179–184.

[4] T. Hirai, J. Yamagishi, and S. Tenpaku, "Utilization of an HMM-based feature generation module in 5 ms segment concatenative speech synthesis," in *Proc. of SSW6*, 2007, pp. 81–84.

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

[6] Z. Ling, L. Qin, H. Lu, Y. Gao, L. Dai, R. Wang, Y. Jiang, Z. Zhao, J. Yang, J. Chen, and G. Hu, "The USTC and iFlytek speech synthesis systems for Blizzard Challenge 2007," in *Proc. of Blizzard Challenge Workshop*, 2007.

[7] Z. Ling and R. Wang, "Minimum unit selection error training for HMM-based unit selection speech synthesis system," in *Proc. of ICASSP*, 2008, pp. 3949–3952.

[8] M. Plumpe, A. Acero, H. Hon, and X. Huang, "HMM-based smoothing for concatenative speech synthesis," in *Proc. of ICSLP*, Sydney (Australia), 1998, pp. 2751–2754.

[9] J. Wouters and M. W. Macon, "Unit fusion for concatenative speech synthesis," in *Proc. of ICSLP*, 2000, pp. 302–305.

[10] V. Pollet and A. Breen, "Synthesis by generation and concatenation of multiform segments," in *Proc. of ICSLP*, 2008, pp. 1825–1828.

[11] T. Okubo, R. Mochizuki, and T. Kobayashi, "Hybrid Voice Conversion of Unit Selection and Generation Using Prosody Dependent HMM," *IEICE Trans. Inf. Syst.*, vol. E89-D, no. 11, pp. 2775–2782, 2006.

[12] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the nitech HMM-based speech synthesis system for the blizzard challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.

[13] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Proc. of MAVEBA*, 2001.

[14] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E90-D 5, pp. 816–824, 2007.

[15] H. Zen, K. Tokuda, T. Masuko, T. Kobayasih, and T. Kitamura, "A Hidden Semi-Markov Model-Based Speech Synthesis System," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[16] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system - HTS-2007 system for the blizzard challenge 2007," in *Proc. of SSW6*, 2007.

[17] J. Yamagishi and T. Kobayashi, "Hidden Semi-Markov model and its speaker adaptation techniques," *IEICE Trans. on Audio, Speech and Language Processing*, vol. 6, 2007.

[18] Y.-J. Wu, W. Guo, and R. Wang, "Minimum generation error criterion for tree-based clustering of context dependent HMMs," in *Proc. of ICSLP*, 2006, pp. 2046–2049.