CONTINUOUS F0 IN THE SOURCE-EXCITATION GENERATION FOR HMM-BASED TTS: DO WE NEED VOICED/UNVOICED CLASSIFICATION?

Javier Latorre, Mark J.F. Gales, Sabine Buchholz, Kate Knill, Masatsune Tamura, Yamato Ohtani, Masami Akamine

1Toshiba Research Europe Ltd., Cambridge Research Laboratory, UK
2Toshiba Corporate Research & Development Center, Japan

ABSTRACT

Most HMM-based TTS systems use a hard voiced/unvoiced classification to produce a discontinuous F0 signal which is used for the generation of the source-excitation. When a mixed source excitation is used, this decision can be based on two different sources of information: the state-specific MSD-prior of the F0 models, and/or the frame-specific features generated by the aperiodicity model. This paper examines the meaning of these variables in the synthesis process, their interaction, and how they affect the perceived quality of the generated speech. The results of several perceptual experiments show that when using mixed excitation, subjects consistently prefer samples with very few or no false unvoiced errors, whereas a reduction in the rate of false voiced errors does not produce any perceptual improvement. This suggests that rather than using any form of hard voiced/unvoiced classification, e.g., the MSD-prior, it is better for synthesis to use a continuous F0 signal and rely on the frame-level soft voiced/unvoiced decision of the aperiodicity model.

Index Terms—Continuous F0, voiced/unvoiced decision, aperiodicity, multi-band mixed excitation, HMM-based synthesis

1. INTRODUCTION

In the source-filter paradigm, the speech signal is decomposed into two parts: a sequence of filter coefficients that represent the vocal tract, and a residual signal that corresponds to the glottal flow. At re-synthesis time, the residual is substituted by a synthetic source excitation signal. In many approaches, this signal consists of a combination of pulses to model the periodic component and white noise to model the aperiodic one. In the primitive vocoder [1], pulse and noise were combined by simply switching between them depending on whether the value of the fundamental frequency (F0) was zero. However, most speech sounds are not purely periodic or aperiodic. Therefore, this approach suffered from a characteristic metallic quality. To reduce this problem, mixed-excitation linear prediction (MELP) was proposed [2]. In MELP, pulse and noise pass first through a rank of N band-pass filters. Then, for each band the outputs are added according to the corresponding aperiodicity value. Recent high quality vocoders such as STRAIGHT [3] extend this idea by using a very fine frequency resolution on the aperiodicity spectrum. This allows re-synthesized speech to be obtained that is almost indistinguishable from the original one.

HMM-based speech synthesis uses the source-filter paradigm to obtain the parameterization of the speech signal from which statistical models are trained. The original implementation of HMM-TTS used pulse/noise switching excitation [4]. As in the corresponding vocoder, the aperiodicity information was embedded within the F0 signal, which was modeled as a discontinuous signal. For this purpose, Multi-space-distribution (MSD) based models were proposed [5]. One problem with this approach was its vocoder-like quality. To improve it, a new source excitation based on MELP was proposed [4], that integrates a 5-band model of the aperiodicity into the synthesis HMMs. The quality of the generated speech was greatly improved, but F0 was still modeled as a discontinuous signal. Therefore, whether a frame is synthesized as unvoiced or as partially voiced depends now on two independent decisions: a state-by-state hard one by the F0 model, and a frame-by-frame soft one by the aperiodicity. As a result, the probability for voiced frames to be wrongly synthesized with pure noise excitation increases.

This paper examines these two devoicing mechanisms and their effect on the perceived quality. The rest of the paper is organized as follows. Section 2 examines the MSD-F0 and aperiodicity models: their characteristics and the voiced/unvoiced decision they produce. Section 3 analyzes the interaction of these two models during synthesis and proposes a new framework that combines their advantages. Section 4 presents several perceptual experiments that analyze the effect on the speech quality of different ways of implementing a hard voiced/unvoiced classification within the proposed framework. Section 5 discusses the results and argues for leaving the voiced/unvoiced decision exclusively to the aperiodicity model. Finally, conclusions are drawn in section 6.

2. VOICED/UNVOICED DECISION

Any hard voiced/unvoiced classification produces two types of errors: false voiced, i.e. setting to voiced frames that should be unvoiced, and false unvoiced, i.e. setting to unvoiced frames that are voiced. Perceptually, false voiced errors produce buzziness, mainly in the higher frequencies, whereas false unvoiced introduce a hoarse quality in the speech signal. In standard HMM-based speech synthesis, this classification is done by the F0 model by means of the MSD-priors. However, even those implementations that consider F0 to be continuous during the training, include some form of hard voiced/unvoiced decision to generate a discontinuous F0 during the synthesis: based on the phone identity [6], based on a separate voicing strength stream [7], or based on the mixture weight [8]. When a HMM-based synthesizer uses a multi-band mixed excitation, this hard decision overlaps with the soft one of the band aperiodicity weights. To understand how these two mechanisms interact, the meaning of each of them will first be examined.
2.1. MSD-prior

Standard HMM-based synthesis uses MSD to model F0 [5]. In MSD, F0 is considered to consist of a unidimensional continuous signal for voiced sounds and a discrete zero-dimensional symbol for unvoiced ones. Therefore, MSD has two distributions: a continuous Gaussian one for the F0 in voiced sounds and a discrete one for the unvoiced symbol. During the training of an MSD model, an observation can be either voiced or unvoiced. Therefore, the state occupancy calculated during the forward-backward alignment has to be fully assigned to the corresponding MSD component. In this way, the MSD-priors represent the probability that a state is voiced. At synthesis time, states are first classified into voiced or unvoiced depending on whether the MSD-prior of the voiced component is above a certain threshold, typically 0.5. Then, the parameter generation algorithm [9] is applied separately to each block of contiguous voiced states.

MSD presents several problems. As shown in [10], it implicitly assumes that the voiced/unvoiced classification of the F0 extraction algorithm is free of errors. As a result, the training becomes very sensitive against errors, especially when they occur at the boundaries of states for which the initial MSD-prior of one of the components is close to zero. In such cases, the other streams cannot compensate the log-likelihood given by that prior. Therefore, the state occupancy probability is ruled by the erroneous F0 observations, deteriorating not only the F0 model but also the duration one. Furthermore, since the F0 model embeds the source excitation, the relevance of segmental context during clustering becomes higher, to the detriment of supra-segmental context. During synthesis, since the F0 of each voiced block is generated in isolation, the long-range correlations of supra-segmental structure become blurred, contributing to the general “flatness” of the HMM-based F0. Moreover, whereas the spectrum and aperiodicity are generated for each frame, the voicing is decided for each state. Often the synthesized spectrum is still voiced during the first frames of a state, but the source excitation is already unvoiced. As a result, hoarseness is introduced in the speech signal.

The training problems can be solved by assuming F0 is a continuous signal, either by making it continuous by means of smoothing and interpolation of the extracted F0 [6],[7], or by assuming that the F0 is generated by two different distributions depending on whether the F0 extraction yields a non-zero value or not [8]. Within the MSD framework, the synthesis problems can be alleviated by reducing the threshold for selecting the states so that a more continuous F0 is created. However, this approach has a risk. The lower the prior of the voiced component, the fewer the samples used to train the Gaussian. Consequently, the less robust the model is against outliers and errors of the F0 extraction algorithm. If anomalous states are used for the F0 generation, their effect is not limited to the frames associated to them, but due to the dynamic features, their effect extends also to the frames associated with normal states. As a result, the generated F0 tends to be more unstable. In this sense, the MSD-prior can be regarded as an indicator of the model robustness.

2.2. Aperiodicity model

The aperiodicity is the proportion of non harmonic components of a signal. In other words, the part of the spectrum that cannot be explained as a weighted sum of the harmonics of the fundamental frequency. The values of the aperiodicity range from 0, for purely periodic, to 1 for purely non-periodic. The real aperiodicity is hard to calculate. However there are multiple ways to obtain a reasonable approximation, for example by comparing the spectrum at the harmonics and inter-harmonic frequencies [3],[11] or by means of band-based normalised correlation coefficients [2],[4]. In standard HMM-based TTS, the aperiodicity spectrum is parameterized by its average value on 5 bands. The band aperiodicity and its dynamic features are modeled by Gaussians and clustered as a separate stream independently from the spectrum and F0. During training, the output probability of the aperiodicity stream is usually ignored to compute the state occupancy probability. At synthesis, the band aperiodicity is generated using the standard algorithm [9]. The generated values are used to control the proportion of pulse and noise in the multi-band mixed excitation. The advantages of a voiced/unvoiced decision based on the aperiodicity are quite clear: whereas the voiced/unvoiced decision of the F0 model is a binary value generated on a state-basis from the prior information of each state by itself, the synthesized band aperiodicity consists of continuous values generated on a frame-basis using static and dynamic features.

3. CONTINUOUS F0 FOR SYNTHESIS

In standard HMM-based TTS with mixed-excitation, both the voiced/unvoiced classification of the F0 model and the soft decision of the aperiodicity model are used. As a result, whereas frames classified as voiced will be excited with a mixture of pulse and noise, frames classified as unvoiced will only be excited with noise. This produces a reduction of the buzziness at the cost of an increment of the hoarseness. To obtain a more balanced excitation, it seems convenient to use only one voiced/unvoiced decision based on the aperiodicity, as in [7].

However, MSD-priors can still be useful to avoid unreliable models in the F0 generation. Fig. 1 depicts a framework for the source excitation generation that aims at getting the best of both models. The main difference with respect to the standard system are the interpolation of the generated F0 and the thresholding on the generated aperiodicity.

A voicing decision based only on the aperiodicity requires a continuous F0 at synthesis. Two ways to obtain this are to train the models on a continuous F0 signal [6],[7]; and to interpolate the generated discontinuous model. In the first case, the discontinuous F0 obtained from the F0 extraction algorithm has to be smoothed, to reduce the effect of errors in the F0 extraction, and interpolated. The way in which this is implemented is not trivial as it affects the quality of the F0 model. The second option is simpler. Moreover, since it uses the same model as the standard approach the duration, spectrum, and general F0 contour are also the same. In this way, perceptual differences will depend only on the aperiodicity of the source excitation, which is the purpose of study of this paper.
The goal of the thresholding module is to transform the soft voiced/unvoiced decision of the aperiodicity into a hard one. This can be expressed in a general way by the following equation

$$\alpha_t = \begin{cases} 1 & \text{if } F(\alpha'_t) \geq \alpha \\ \alpha'_t & \text{otherwise} \end{cases}$$

where $F$ and $\alpha$ are any arbitrary function and threshold.

4. EXPERIMENTS

Several subjective experiments were designed to understand the effect of the different devoicing mechanisms.

4.1. Experimental set-up

The HMMs were trained on 4369 sentences (approximately 4.5 hours of speech) recorded from a single American English female speaker and sampled at 16kHz. The models were 5-state left-to-right without skip. The observation vector consists of the static and delta of 40 LSP coefficients, F0, and 5 aperiodicity bands (0 to 1kHz, 1 to 2.25kHz, 2.25 to 4kHz, 4 to 6kHz, and 6 to 8kHz). The spectrum was obtained with a pitch synchronous analysis, and the aperiodicity with PSHF [11]. Models were trained using HTS [12].

The experiment consisted of several pair test evaluations conducted on Amazon Mechanical Turk (AMT). For each evaluation 50 pair stimuli were randomly selected from a pool of 224 possible ones. Each pair stimulus was duplicated with reversed order and evaluated by 4 different subjects, i.e., each pair stimuli received a total of 8 judgments. Only subjects registered in the US were allowed to participate. In AMT there is no perfect control on the number of different subjects that participate in an experiment. However, the evaluations were set in such a way that at least 13 different subjects took part in each one. In practice, the number of different subjects per evaluation ranged from 19 to 71.

4.2. Modifying the threshold on the MSD-Prior $\beta$

The first experiment consisted of modifying the threshold on the MSD-prior $\beta$. For this experiment, the standard framework was used, i.e., without interpolation and thresholding. Utterances synthesized with different MSD-prior thresholds were compared against those synthesized with the default $\beta = 0.5$. As table 1 shows, whereas for $\beta > 0.5$ the quality degrades quickly due to the higher number of false unvoiced, i.e., higher hoarseness, reducing $\beta$ does not seem to have any negative effect. The optimum value seems to be around $\beta = 0.2$. For $\beta = 0.1$ the quality does not decrease but it does not improve either. This is probably because the higher instability on the generated F0 balances out the advantage of reduced hoarseness.

<table>
<thead>
<tr>
<th>MSD-prior ($\beta$)</th>
<th>System Preference Standard</th>
<th>System Preference Non-standard</th>
<th>No Pref.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>31.7%</td>
<td>33.8%</td>
<td>34.5%</td>
<td>0.34</td>
</tr>
<tr>
<td>0.2</td>
<td>29.7%</td>
<td>37.4%</td>
<td>32.7%</td>
<td>0.07</td>
</tr>
<tr>
<td>0.7</td>
<td>40.1%</td>
<td>26.9%</td>
<td>33.0%</td>
<td>0.004</td>
</tr>
<tr>
<td>0.9</td>
<td>53.3%</td>
<td>19.2%</td>
<td>27.4%</td>
<td>&lt;1e-6</td>
</tr>
</tbody>
</table>

4.3. Threshold on the aperiodicity

The purpose of the next experiments was to test whether the proposed framework could yield better results. In other words, whether the voiced/unvoiced decision can rely entirely on the aperiodicity; and whether some form of hard decision would help. To avoid variability, $\beta$ was fixed to 0.5 across all the experiments. First, the interpolated F0 without any thresholding ($\alpha = 1$) was compared with the non-interpolated one. The results yield a significant preference ($p < 0.024$) for the interpolated F0 of 34.7% vs 24.7%, with 40.6% of the answers being undecided. Next, two different functions to obtain a threshold on the aperiodicity were tested: the average of the 5-band aperiodicity vector of a frame, and the value of the lower aperiodicity band. Figure 2 shows the expected misclassification for the average and the lower band aperiodicity over the training data. The crossing point for the average band aperiodicity is at 0.58 and for the first band aperiodicity at 0.029. Speech samples generated with threshold function at several threshold values were compared with those generated without a threshold ($\alpha = 1$). The results are shown in table 2. When the threshold is applied on the lower aperiodicity band no difference in the preference is detected, even though with a threshold $\alpha_{avg} = 0.5$ the probability of false voiced decreases from 100% to less than 37%. On the other hand, when the threshold is applied on the average aperiodicity the quality decreases rapidly for $\alpha_{avg} < 0.8$. Around this value there seems to be some improvement, though not statistically significant. Interestingly, for $\alpha_{avg} > 0.8$ the probability of false unvoiced decreases rapidly so that for $\alpha_{avg} = 0.86$ it is less than 0.1%. For this threshold the probability of false voiced decreases to 42%. Yet, subjects did not perceive any difference with respect to $\alpha = 1$.

<table>
<thead>
<tr>
<th>Aperiodicity threshold ($\alpha$)</th>
<th>System preference No threshold</th>
<th>System preference Threshold</th>
<th>No Pref.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{avg} = 0.5$</td>
<td>29.6%</td>
<td>28.9%</td>
<td>41.6%</td>
<td>0.44</td>
</tr>
<tr>
<td>$\alpha_{avg} = 0.7$</td>
<td>31.8%</td>
<td>30.3%</td>
<td>37.9%</td>
<td>0.38</td>
</tr>
<tr>
<td>$\alpha_{avg} = 0.9$</td>
<td>30.6%</td>
<td>28.9%</td>
<td>40.5%</td>
<td>0.36</td>
</tr>
<tr>
<td>$\alpha_{avg}(\beta) = 0.6$</td>
<td>55.6%</td>
<td>19.7%</td>
<td>24.7%</td>
<td>&lt;1e-10</td>
</tr>
<tr>
<td>$\alpha_{avg}(\beta) = 0.7$</td>
<td>44.6%</td>
<td>25.5%</td>
<td>29.9%</td>
<td>&lt;1e-4</td>
</tr>
<tr>
<td>$\alpha_{avg}(\beta) = 0.8$</td>
<td>25.8%</td>
<td>31.0%</td>
<td>43.2%</td>
<td>0.16</td>
</tr>
<tr>
<td>$\alpha_{avg}(\beta) = 0.86$</td>
<td>26.8%</td>
<td>24.1%</td>
<td>49.1%</td>
<td>0.29</td>
</tr>
</tbody>
</table>
4.4. Interpolation + threshold vs MSD-prior

In a final evaluation, samples generated with discontinuous F0 and the optimized MSD-prior $\beta = 0.2$, were compared with those generated with $\beta = 0.5$ and interpolated F0 at different values of the aperiodicity threshold ($\alpha$). As shown in table 3, no significant difference was found. In this sense, it is interesting to notice how a trivial linear interpolation of the F0 generated with a standard $\beta$, equals or exceeds the performance obtained with an optimized one.

Table 3. Preference for different forms of voiced/unvoiced decision.

<table>
<thead>
<tr>
<th>Aperiodicity threshold ($\alpha$)</th>
<th>System Preference</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>None ($\alpha = 1$)</td>
<td>Disc. F0</td>
<td>0.34</td>
</tr>
<tr>
<td>$\alpha_{avg}(\beta) = 0.86$</td>
<td>Interp. F0</td>
<td>0.48</td>
</tr>
<tr>
<td>$\alpha_{avg}(\beta) = 0.80$</td>
<td>No pref.</td>
<td>0.13</td>
</tr>
</tbody>
</table>

5. DISCUSSION

In this paper different ways to obtain a voiced/unvoiced decision for HMM-based synthesis have been explored. The experimental results show that, when using MELP-like excitation, there is a strong asymmetry in the way false unvoiced and false voiced errors are perceived. Whereas the first rapidly decrease the speech quality, the second are mostly unperceived. Such asymmetry seems to indicate that the soft weights of the band aperiodicity are enough to eliminate most of the buzziness from unvoiced frames. Therefore, no additional hard voiced/unvoiced decision is required. Accepting that the voicing decision depends exclusively on the aperiodicity relieves F0 from any "responsibility" about the voiced/unvoiced classification. This allows F0 to be modeled as a standard continuous signal. Some preliminary results using continuous F0 both for training and synthesis show that this could yield a better speech quality with a significant preference score of 66.3% vs. 33.7% over the standard discontinuous F0 for a Japanese HMM-based TTS system. In this way, in addition to producing better F0 and duration models [6],[7],[10], it might be also possible to obtain more expressive and natural F0 contours by facilitating the integration of F0 models at higher supraglottal levels.

Of course, another possible explanation for the perceptual asymmetry is that most subjects recruited via AMT could not really perceive the buzziness due to the distortion introduced in high frequencies by standard home audio equipments. It is logical to expect that in a controlled experiment with high-fidelity head-phones and audiocards subjects would be more annoyed by the false voiced errors. Further experiments are needed to clarify this point. Nonetheless, it should be noticed that the listening conditions of AMT are closer to those of most real-life applications.

6. CONCLUSIONS

This paper analyzes the effect of applying a hard voiced/unvoiced decision in HMM-based speech synthesis when a MELP-like source excitation is used. Three forms of hard decision were considered: a state-by-state one based on the MSD-prior, and two frame-by-frame, one based on the average band aperiodicity and another based on the value of the lower aperiodicity band. The results indicate a strong asymmetry in the way subjects judge the harshness produced by false unvoiced errors and the buzziness produced by false voiced ones. Regardless of which voiced/unvoiced classifier was used, subjects consistently preferred the samples generated from configurations with a minimal or zero probability of false unvoiced errors, so that few or no voiced sounds are synthesized using only noise excitation. On the other hand, differences in the number of false voiced errors produced little or no effect on the perceived quality. This suggests that it is preferable for the F0 model to produce a continuous signal and rely on the frame-based multi-band mixed excitation to reduce the buzziness to acceptable levels.

7. REFERENCES