PUB Entry in the Blizzard Challenge 2011

Marius Cotescu

Department of Applied Electronics and Information Engineering, "Politehnica" University of Bucharest, Bucharest, Romania

mcotescu@lpsv.pub.ro

Abstract

The paper presents the entry in this year's Blizzard Challenge of the Politehnica University of Bucharest. We present a parametric speech synthesis system based on HTS, which tried to achieve two important goals: gain better control over the vocal tract filter, and allow greater variability for the excitation source features by separating, as much as possible, the two processes. We proposed the spectral tilt as a feature of both voiced and unvoiced excitation that can be easily and reliably estimated and extracted from the smoothed spectrum, leaving more consistent data for the vocal tract model. We also engaged the problem of modelling the STRAIGHT aperiodicity coefficients in a new manner, which provides more details to synthetic speech. It was the first entry from the laboratory, and unfortunately both the limited experience and the scarce human resource deployed had decisively influenced the results.

Index Terms: speech synthesis, spectral tilt, glottal flow, HMM

1. Introduction

Parametric speech synthesis systems based on the HMM framework have always had the advantage of intelligibility over concatenative systems, while naturalness, despite continuous improvements [1], [2], [3], [4], remains its weakest point. One of the constantly reported shortcomings of HMM-based TTS systems is the lack of the "small imperfections" characteristic to natural speech.

One way of addressing this problem is the use of HMM-driven unit selection systems that are able to deliver high naturalness [5], as well as high intelligibility [6]. Another way seems to be the careful refinement of the statistical models [3], which has shown that purely parametric systems can produce very competitive systems.

In the source-filter paradigm of speech production [7], the vocal tract is viewed a being a passive component, that brings an almost negligible contribution to the energy of the spoken signal. Further more, an attempt is made to model the spectral tilt of the signal, which is strongly correlated to the glottal pulse shape [8], as an independent feature of the spoken signal. Previous attempts have been made to use a glottal flow model to excite HMM-based synthesis systems [9], but robustness problems in parameter extraction were reported.

A cepstral approach is introduced to model the STRAIGHT aperiodicity coefficients [10], in order to achieve finer details of the generated excitation signal.

Section 2 presents the signal processing techniques deployed to remove excitation signal information from the vocal tract filter data, and our view on modelling the aperiodicity coefficients. Section 3 briefly describes the implementation of our modelling choices into an HTS-based system, and our labelling scheme. Section 4 presents the results and Section 5 concludes the paper.

2. Feature Extraction

To address the lack of variability in synthetic speech, we first need to identify its sources in natural speech. The source-filter model [7] describes two separate processes that take place during speech production: producing the excitation signal, and then re-shaping its spectral structure by passing through the vocal tract. The vocal tract shape is responsible for encoding the lexical information, and needs to vary its configuration at a slower rate to allow the listeners time to decode the message. The excitation signal, on the other hand, is conveying only broad lexical information, and is allowed to vary its parameters more freely. The small contribution to the lexical information, does not mean that it does not contribute to the conveyed information. Quite the contrary, it is often associated with conveying emotions and subtle changes in the meaning. The small and rapid perturbations in the excitation parameters are also responsible for the naturalness of the uttered sound.

In [11] we have proven that synthetic speech quality can be improved by separating the energy and spectral tilt information from the vocal tract model. This allowed the vocal tract models to be more accurate, while allowing the other spectral feature to vary independently, thus permitting the system to generate a greater range of possible acoustical realisations.

2.1. Spectral Structure of the Glottal Pulse

In [11] the spectral tilt was defined as the regression line fitted over a frame's amplitude spectrum, expressed in decibels, over a logarithmic frequency scale. This definition — although proving to give results — is not in accordance with the spectral structure of the glottal pulse.

We have chosen to use the Liljencrants-Fant (LF) glottal flow model, [12], in order to analyse the spectral structure of the glottal pulse, and thus improve the working definition of the spectral tilt. The time-domain expression of the LF model is given in Equation 1. The first branch of the model describes the flow during the opening phase of the glottal movement, which starts at $t = 0$ and ends at the moment of maximum excitation, t_e. The second branch defines the behaviour during the closing stage. The time constant t_a is defined as the duration from t_e to the moment when a tangent at the exponential in $t = t_e$ reaches 0. The other two parameters are the maximum excitation E_e, and the instant of maximum airflow t_p.

$$e(t) = \begin{cases} E_0 \cdot e^{\alpha t} \cdot \sin(\omega_g t), & 0 \leq t \leq t_e \\ -\frac{E_e}{\epsilon t_a} \cdot \left(e^{-\epsilon(t-t_e)} - e^{-\epsilon(T-t_e)} \right), & t_e < t \leq T \end{cases} \quad (1)$$

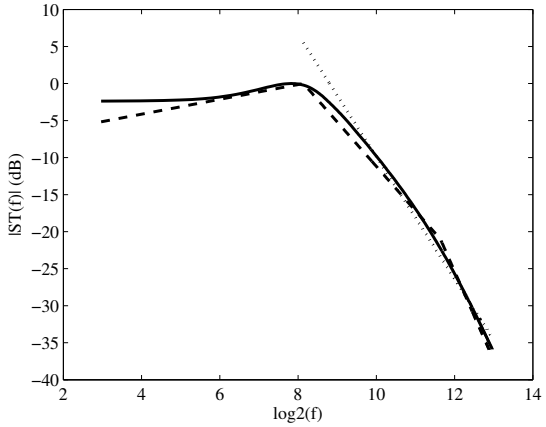

Figure 1: *Spectrum of the LF model (solid line), its stylisation (dashed line), and its approximation by a regression line in the log-frequency domain (dotted line). [$f_0 = 150Hz, t_p = 2.8ms, t_e = 4ms, t_a = 0.05ms$]*

where $\omega_g = \frac{1}{\pi t_p}$. The values of E_0, α, and ϵ can be determined using the methods described in [12].

The stylisation of the LF glottal flow model spectrum [8] is given in Fig. 1. After the maximum represented by the glottal formant F_g, the amplitude decreases at a rate of 6 dB/octave, until the cut-off frequency F_c of the low-pass filter, introduced by the return phase. The two combined filters give a slope of -12 dB/octave.

By approximating the spectral shape of the glottal pulse by a single line segment, we were able to remove some of the aspects of the excitation signal from the vocal tract model. The fact that the spectral tilt can also be defined for unvoiced sounds, ensures the generality of the measure. Thus, the chances of artefacts appearing in the generated utterances at voiced-unvoiced transitions is significantly reduced.

However, this is not a very good approximation of the spectral shape of the glottal pulse. One solution would be to determine the temporal parameters of the LF model for each frame, and then subtract the corresponding spectrum from the observed frame. But, by doing this the advantage of a smooth transition of the extracted features between voiced and unvoiced segments would be lost.

Instead, we focused on maintaining the approach, but minimizing the approximation error. To achieve this, we tried to find the optimum value for a of the frequency warping function

$$\beta_a(\omega) = \tan^{-1}\left(\frac{1 - a^2\sin(\omega)}{1 + a^2\cos(\omega) - 2a}\right) \qquad (2)$$

over which the spectrum would be better fitted by a linear model. To do this, we have generated synthetic glottal pulses covering a large enough domain in the parameter space that were used to evaluate the different warping factors. The range of the parameters is presented in Table 1.

For each value of the warping factor $a \in [0.38, 0.58]$, the mean absolute error between the linear approximation of the spectrum, and it's actual realisation, expressed in dB, was computed. Fig. 2 shows a plot of the MSE versus the different values of alpha. The dashed line represents the baseline error obtained by fitting the spectrum with two linear segments (from

Table 1: *Range of the LF model parameters used in evaluations.*

Parmeter	Units	Range
f_0	Hz	$[150, 350]$
t_p	% of t_e	$[70, 80]$
t_e	% of T_0	$[60, 70]$
t_a	ms	$[0.05, 0.3]$

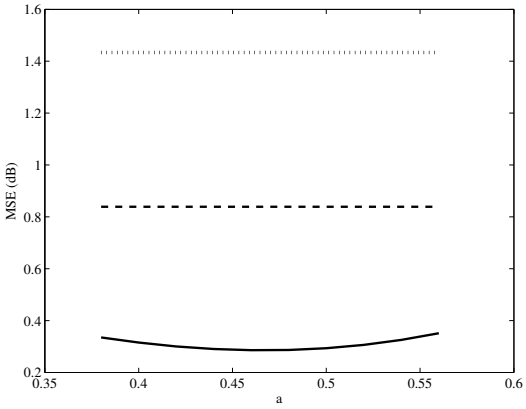

Figure 2: *MSE vs warping factor.*

F_g to F_c, and from F_c to the Nyquist frequency) on the logarithmic scale. The dotted line represents the baseline obtained by fitting the spectrum with only one linear segment between F_g and the Nyquist frequency over the logarithmic scale. The MSE shows a minimum of 0.285 dB for $a = 0.46$, which for the sampling frequency of 16 kHz, falls somewhere between the Mel ($a = 0.42$) and Bark ($a = 0.57$) frequency scales.

As the global spectral tilt is actually given by the lengths of the two linear segments on the logarithmic scale, i.e. the value of F_c, we would also desire a frequency scale that provides a strong correlation between t_a and the spectral tilt. Fig. 3 shows the correlation coefficient between the measured spectral tilt and the value of t_a as a function of a. The mean value of the measured spectral tilt as a function of t_a and a is shown in Fig. 4.

Because of the very small variation in both the MSE value, and the correlation coefficient, a preliminary listening test was run to choose between the Mel and the Bark frequency scales. Two systems were trained, using the first 1000 sentences in the corpus, and the naturalness of the two voices was evaluated, showing a slight preference for the Bark scale.

We define the spectral tilt of a frame, m, as the slope of the regression line fitted over the amplitude spectrum expressed in dB, starting from $2f_0$ to $f_s/2$, over the Bark scale. The log-energy of the frame $\log E$, is defined as the intercept of the regression line. Notice that this stylisation ignores the position of the glottal formant. We start the fitting at $2f_0$ in order to make sure that the glottal formant will not affect the slope. For unvoiced frames, f_0 is set to 150 Hz. The gross representation of the excitation signal $ST(\omega)$, can then be written as

$$\log ST(f) = \log E + m \cdot \textit{Bark}(f), \qquad (3)$$

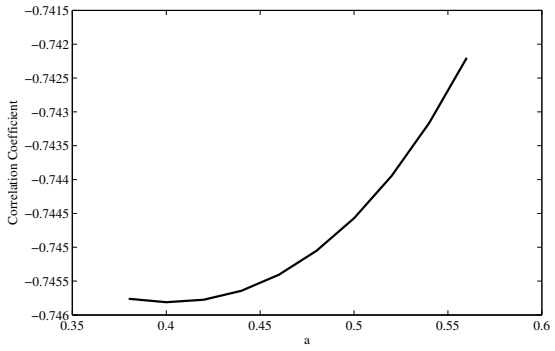

Figure 3: *Correlation coefficient between the measured spectral tilt and the value of t_a.*

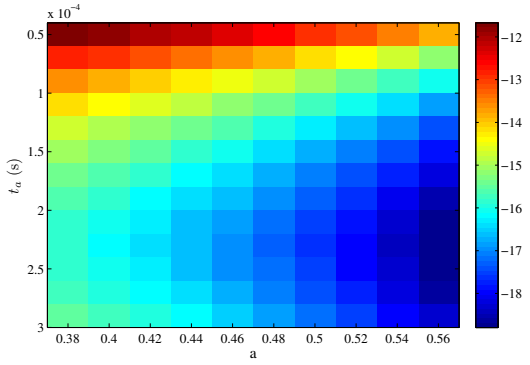

Figure 4: *Mean value of the measured spectral tilt as a function of t_a and a.*

where

$$Bark(f) = 13\arctan(0.00076 \cdot f) + 3.5\arctan\left((\frac{f}{7500})^2\right) \quad (4)$$

2.2. Extracting the Vocal Tract Filter

The spectrum of an analysis frame of vocal signal $S(\omega)$ is determined by a number of contributions along the vocal tract length. First, there is the general structure $ST(\omega)$ given by the excitation signal, superimposed on a harmonic structure $H_a(\omega)$ (if the sound is voiced), or noise (in the case of unvoiced sounds). The excitation signal is then filtered by the vocal tract ($VT(\omega)$), and radiated from the lips ($U(\omega)$).

$$S(\omega) = H_a(\omega) \cdot ST(\omega) \cdot VT(\omega) \cdot U(\omega). \quad (5)$$

The harmonic structure is removed by STRAIGHT analysis [13], and it's usual to include the radiation filter in the excitation model. Thus, the spectrum produced by STRAIGHT analysis can be written in a simpler manner, as

$$S(\omega) = ST(\omega) \cdot VT(\omega). \quad (6)$$

By using the stylisation of the excitation signal introduced in Equation 3, the vocal tract transfer function can be approximated by

$$\log(VT(\omega)) = \log(S(\omega)) - (\log E + m \cdot \beta_a(\omega)). \quad (7)$$

The spectrum obtained from Equation 7, is then processed and 39 generalised cepstral coefficients [14] ($\gamma = 0.33$), warped on the Bark scale ($a = 0.57$), without the 0^{th} coefficient, are extracted.

2.3. Modelling the Aperiodicity Coefficients

The aperiodicity (AP) coefficients, introduced in [10] as the ratio between the STRAIGHT spectral envelope extracted using the harmonic peaks and the one extracted using inter-harmonic valleys, are a measure of the jitter, shimmer, and Harmonic-to-Noise Ratio. The classical way of coding the AP coefficients is by using their mean value over five frequency sub-bands (0–1 kHz, 1–2 kHz, 2–4 kHz, 4–6 kHz, and 6–8 kHz). A perceptually motivated approach, using the Bark critical sub-bands, was proposed in one of the entries in the previous Blizzard Challenge [3] showing that a more refined coding can improve the naturalness of the synthetic voice.

Our system, uses the method presented in [11]. The AP coefficients are treated as samples of an amplitude spectrum,

$$AP[n] = P_{AP}(\pi \cdot \frac{n}{N} \cdot f_S), \quad (8)$$

from which cepstral coefficients can be extracted. Our studies showed that by replacing the 5 sub-band mean values with 15-order cepstral analysis produces a noticeable increase in the quality of the synthetic voice.

3. Synthesis System

The synthesis system is based on the demo provided with version 2.1.1 of the HTS toolkit. It uses 5-state trajectory-HSMM models, with explicit state duration modelling. The missing values for the pitch in unvoiced frames are modelled using the MSD framework.

3.1. Structure of the Acoustic Models

The observation vectors include information about the vocal tract, pitch, energy, spectral tilt, and the AP coefficients, structured in 8 streams:

- **Vocal Tract** information is coded using one stream, containing the 39 generalised cepstral coefficients, their Δ, and Δ^2 coefficients.
- **Pitch** information is coded using the current value extracted by STRAIGHT analysis, expressed in Mels, and its first and second derivatives, in three separate MSD streams.
- **AP coefficients** are coded using 16 cepstral coefficients, and their first and second derivative, in one stream.
- **Energy and Spectral Tilt** are coded in one stream, due to the strong link between the two, introduced by the way in which they are defined. Alongside their current values, the first and second derivatives are included in two additional streams.

3.2. Labelling

The labelling scheme generally follows labelling scheme introduced in [15], with the exception of the phonetic and syllable levels, which were slightly affected by the use of lessemes as the base labels.

At the phonetic level, the phoneme was replaced by the base lesseme. At the syllable level, the stress information usually

provided by the lexicon was extracted from the lesseme stress value. Additional fields were inserted at the syllable level to label the additional pitch level and inflection information provided by the vowel lessemes. For consonant sounds, the playability information was kept in the phonetic context.

The question set used in the clustering stage is derived from the questions provided with the HTS demo. The following questions were added:

- is the value of {preceding, current, succeeding} syllable pitch level equal to $k \in \{1, 2, 3\}$?
- is the value of {preceding, current, succeeding} syllable pitch level less than or equal to $k \in \{2, 3\}$?
- is the value of {preceding, current, succeeding} syllable inflection type equal to {fN, fU, fW, fC, fD, fS}?
- what is the playability of the {preceding, current, succeeding} consonant?

3.3. Waveform Synthesis

For each frame, the system produces cepstral coefficients for the vocal tract and AP models, a value for the frame's pitch, and the energy and spectral tilt information. After reconstructing the vocal tract transfer function, and the AP coefficients, the STRAIGHT spectrum is reconstructed using Equations 3 and 6. The smooth spectrum, AP coefficients, and pitch contour are fed to the STRAIGHT vocoder to reconstruct the waveform.

Unfortunately, due to a bug in the spectrum reconstruction code, instead of using the Bark frequency scale, the logarithmic one was used instead. Because the team had only one person, the final version was ready extremely late in the schedule, and exhaustion had set in, no effort was put into finding and fixing the bug. Even though the preliminary voices, built for evaluating the frequency scales, were significantly better than the final one.

The bug was fixed *after* the results were published, and the new utterances can be downloaded from ftp://lpsv.pub.ro. Too late, though.

4. Results

This year's Blizzard Challenge was organised into one hub task, EH1, and one spoke task, ES1. The hub task consisted of building a general voice using the entire database. The spoke task consisted of building a voice designed for reading names and addresses in US format. The EH1 evaluation consisted of the standard naturalness, similarity and SUS intelligibility tests, while the ES1 evaluation focused only on the intelligibility test. We have entered the same voice in both tasks. No particular emphasis or effort was put into designing the voice for the address reading task.

Our system was designated by the letter I in the listening tests. The fact that a mean value of about -8 dB/octave for the spectral tilt on the logarithmic scale for voiced segments was replaced by a mean value of about -3 had disastrous effects on the naturalness of the synthetic utterances. The voice is too "bright", and too "buzzy". The system was placed last in all naturalness and similarity tests.

As far as the intelligibility tests are concerned, the results are a little bit better. In the SUS test, the Word Error Rate is about the same as the poor performing systems, and the differences between our system and the well-performing ones are statistically significant at a 1% level.

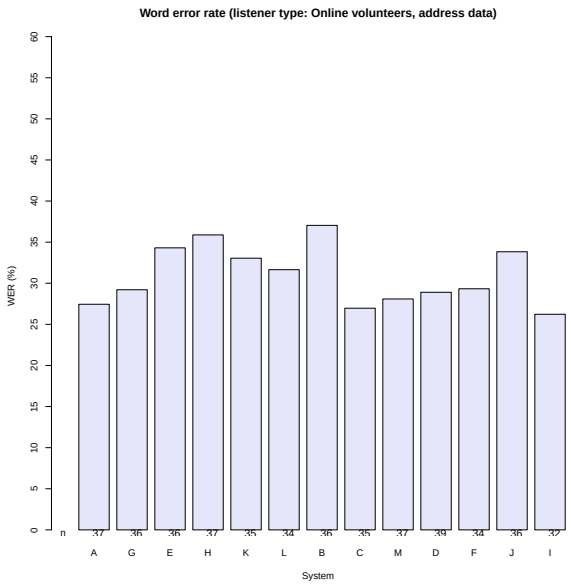

Figure 5: *WER results in the address task for online volunteers.*

In the case of the address data, the system achieved the lowest WER in the online volunteer scenario (Fig. 5), slightly outperforming even natural speech. This might be the result of an over-accentuated Lombard effect, which counted in the scenario of more noisy listening conditions and poorer quality of the equipment. The limited number of data points compels us to take a more reserved position, although the most intelligible system in the 2010 Blizzard Challenge speech in noise task [4] reported, as the main techniques, increasing the spectral tilt and the formant contrast.

5. Conclusions

The paper presents the PUB entry in the 2011 Blizzard Challenge, which introduced improvements to the vocal tract and excitation source coding and models. Although a previous study [11] showed that the deployed methods should have increased the quality of the voice, an unfortunate bug in the system prevented its correct placement in the evaluation. The mistake has produced one interesting result, though: the system had the lowest WER in the ES1 task evaluation, most probably due to the exaggerated Lombard effect produced by the unnaturally high spectral tilt values.

6. Acknowledgements

The author is supported by POSDRU project POSDRU/6/1.5/S/16. The system was mainly developed while the author was visiting the Speech Communication and Technology (CTT) Group at the Royal Institute of Technology (KTH) in Stockholm, Sweden.

The author would like to thank Kornel Laskowski, at CTT, for the help provided with mapping the lesseme to the CMU phoneme set, and the CTT department for kindly providing access to their infrastructure.

7. References

[1] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," Pittsburgh, Jun. 2004.

[2] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," vol. II, Pittsburgh, Oct. 2004, pp. 1397–1400.

[3] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS System for Blizzard Challenge," Kansai Science City, Japan, Sep. 2010.

[4] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM Speech Synthesis Entry for Blizzard Challenge 2010," Kansai Science City, Japan, Sep. 2010.

[5] Y. Jiang, Z.-H. Ling, M. Lei, C.-C. Wang, L. Heng, Y. Hu, L.-R. Dai, and R.-H. Wang, "The USTC System for Blizzard Challenge 2010," Kansai Science City, Japan, Sep. 2010.

[6] Y. Qian, Z.-J. Yan, Y.-J. Wu, F. K. Soong, G. Zhang, and L. Wang, "An HMM Trajectory Tiling (HTT) Approach to High Quality TTS - Microsoft Entry to Blizzard Challenge 2010," Kansai Science City, Japan, Sep. 2010.

[7] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton Co, 1960.

[8] B. Doval, C. D'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[9] J. S. Andersson, J. P. Cabral, L. Badino, J. Yamagishi, and R. A. Clark, "Glottal source and prosodic prominence modelling in HMM-based speech synthesis for the Blizzard Challenge 2009," in *The Blizzard Challenge 2009*, Edinburgh, U.K., Sep. 2009.

[10] 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. MAVEBA 2001*, Firentze, Italy, Sep. 2001.

[11] M. Cotescu and I. Gavat, "Sources of increased variability in HMM synthetic voices," in *Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on*, Brasov, Romania, May 2011, pp. 1–6.

[12] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[13] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[14] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis – a unified approach to speech spectral estimation," in *Proc. ICSLP-94*, 1994, pp. 1043–1046.

[15] K. Tokuda, H. Zen, and B. A.W., "An HMM-based speech synthesis system applied to English," in *Proc. of 2002 IEEE SSW*, Sep. 2002.