SIMPLE METHODS FOR IMPROVING SPEAKER-SIMILARITY OF HMM-BASED SPEECH SYNTHESIS

Junichi Yamagishi, Simon King

The Centre for Speech Technology Research,
University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom
jyamagis@inf.ed.ac.uk

ABSTRACT

In this paper we revisit some basic configuration choices of HMM-based speech synthesis, such as waveform sampling rate, auditory frequency warping scale and the logarithmic scaling of F_0, with the aim of improving speaker similarity which is an acknowledged weakness of current HMM-based speech synthesisers. All of the techniques investigated are simple but, as we demonstrate using perceptual tests, can make substantial differences to the quality of the synthetic speech. Contrary to common practice in automatic speech recognition, higher waveform sampling rates can offer enhanced feature extraction and improved speaker similarity for speech synthesis. In addition, a generalized logarithmic transform of F_0 results in larger intra-utterance variance of F_0 trajectories and hence more dynamic and natural-sounding prosody.

Index Terms— HMM, speech synthesis, HTS, TTS

1. INTRODUCTION

Statistical parametric speech synthesis based on hidden Markov models (HMMs) has become established and widely studied. It has the ability to generate natural-sounding synthetic speech [1] and in recent years, some HMM-based speech synthesis systems have reached performance levels comparable to state-of-the-art unit selection systems [2] in terms of naturalness and intelligibility. However, relatively poor perceived "speaker similarity" is one of the most common shortcomings of such systems [3].

One possible reason for may be the use of a vocoder, which can result in buzziness. Another reason may be that the statistical modelling can lead to a muffled sound, presumably due to the process of averaging many short-term spectra, which removes important detail. In addition to these intrinsic reasons, we hypothesize that there are also extrinsic problems: some basic configuration choices in HMM synthesis have been simply taken from different fields such as speech coding, automatic speech recognition (ASR) and unit selection synthesis. For instance, 16 kHz is generally regarded as a sufficiently-high waveform sample rate for speech recognition and synthesis because speech at this sample rate is intelligible to human listeners. However speech waveforms sampled at 16 kHz still sound slightly muffled when compared to higher sample rate. HMM synthesis has already demonstrated levels of intelligibility indistinguishable from natural speech [2], but high-quality TTS needs also to achieve naturalness and speaker similarity. Therefore we decided to revisit these apparently basic issues and discover whether current configurations are satisfactory, especially with regard to speaker similarity. As the sampling rate increase, the differences between different auditory frequency scales such as the Mel and Bark scales [4] implemented using a first-order all-pass function also become larger. Therefore we also included a variety of different auditory scales in our experiments using higher sampling rates.

Lower speaker similarity may also be caused by problems with the excitation modeling as well as spectral modeling; advanced excitation modeling methods such as mixed excitation [5] have already been proposed to address this. Fundamental frequency (F_0), the main parameter of the excitation signal, is modelled on a logarithmic scale by the HMMs. This is motivated by the Fujisaki model [6] and the fact that log F_0 has a more Gaussian distribution than the raw value. However we have found, over the course of building thousands of HMM-based synthetic voices [7], generated F_0 trajectories tend to be relatively monotonic and lacking the vivid prosody of natural speech, especially for female voices. A simple logarithmic transform may give excessive compression at higher values of F_0. Therefore we employed a generalized logarithmic transformation (also known as a box-cox transform [8]) and used a data-driven maximum-likelihood estimator to set the parameter of this transform that controls the degree of compression. Although none of the techniques above are in themselves new, we found that they can have substantial effects on HMM-based speech synthesis.

This paper is organized as follows. Section 2 gives an overview of the first-order all-pass filter used for mel-cepstral analysis. The use of Bark and ERB scales in the filter is given in Section 3. Section 4 describes the data-driven transformation of fundamental frequency using the generalized logarithmic transformation. System details and experimental results are given in Section 5 and Section 6 concludes the paper by briefly summarizing our findings.

2. THE FIRST-ORDER ALL-PASS FREQUENCY-WARPING FUNCTION

In mel-cepstral analysis [9], the vocal tract transfer function $H(z)$ is modelled by M-th order mel-cepstral coefficients $\boldsymbol{c} = [c(0), \ldots, c(M)]^\top$ as follows:

$$H(z) = \exp \boldsymbol{c}^\top \tilde{\boldsymbol{z}} = \exp \sum_{m=0}^{M} c(m) \tilde{z}^{-m}, \tag{1}$$

where $\tilde{\boldsymbol{z}} = [1, \tilde{z}^{-1}, \ldots, \tilde{z}^{-M}]^\top$. $\tilde{z}^{-1}$ is defined by a first-order all-pass (bilinear) function

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \tag{2}$$

and the warped frequency scale $\beta(\omega)$ is given as its phase response:

$$\beta(\omega) = \tan^{-1} \frac{(1 - \alpha^2) \sin \omega}{(1 + \alpha^2) \cos \omega - 2\alpha}. \tag{3}$$

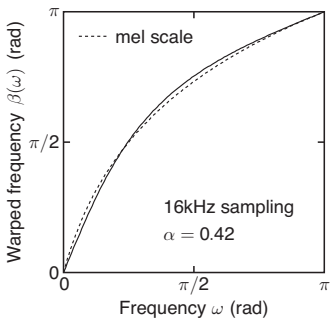

Fig. 1. Frequency warping using the all-pass function. At a sampling rate of 16 kHz, $\alpha = 0.42$ provides a good approximation to the mel scale.

The phase response $\beta(\omega)$ gives a good approximation to an auditory frequency scale with an appropriate choice of α.

An example of frequency warping is shown in Fig. 1. where it can be seen that, when the sampling rate is 16 kHz, the phase response $\beta(\omega)$ provides a good approximation to the mel scale for $\alpha = 0.42$. The choice of α depends on the sampling rate used and the auditory scale desired. The next section describes how to determine this parameter for a variety of auditory scales.

3. THE BARK AND ERB SCALES USING THE FIRST-ORDER ALL-PASS FUNCTION

In HMM-based speech synthesis, the mel scale is widely used. For instance, Tokuda *et al.* provide appropriate α values for the mel scale for speech sampling rates from 8kHz to 22.05kHz [10]. In addition to the mel scale, the Bark and equivalent rectangular bandwidth (ERB) scales [11] are also well-known auditory scales. In [12], Smith and Abel define the optimal α (in a least-squares sense) for each scale as follows:

$$\alpha_{\mathrm{Bark}} = 0.8517\sqrt{\arctan(0.06583\, f_s)} - 0.1916 \quad (4)$$

$$\alpha_{\mathrm{ERB}} = 0.5941\sqrt{\arctan(0.1418\, f_s)} + 0.03237 \quad (5)$$

where f_s is the waveform sampling rate. However, note that the error between the true ERB scale and all-pass scale approximated by α_{ERB} is three times larger than the error for the Bark scale using α_{Bark} [12]. Note also that as sampling rates become higher, the accuracy of approximation using the all-pass filter becomes worse for both scales.

4. DATA-DRIVEN TRANSFORMATION OF FUNDAMENTAL FREQUENCY

4.1. The generalized logarithmic (box-cox) transform

The generalized logarithmic transform function G of fundamental frequency F_0 is defined as follows:

$$G(F_0, \lambda) = \begin{cases} \log F_0 & \text{if } \lambda = 0 \\ (F_0^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \end{cases} \quad (6)$$

where λ is a parameter for determining how much the transform "compresses" F_0. We refer the transform of F_0 via G as "generalized $\log F_0$".

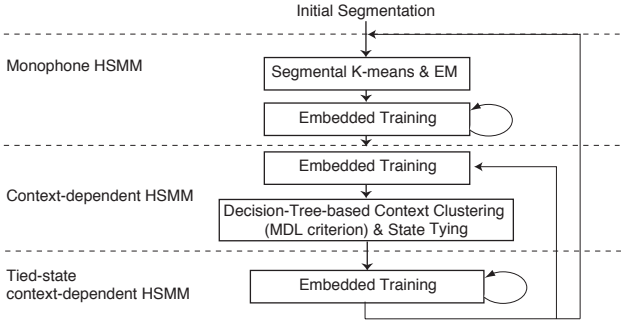

Fig. 2. Overview of HMM training stages for HTS voice building.

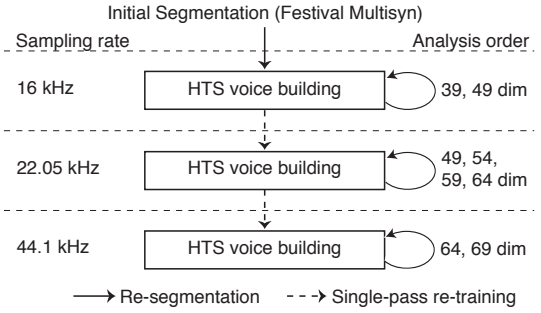

Fig. 3. Bootstrap voice building. Training starts by building models for lower sampling rate speech with a lower analysis order and gradually increases the analysis order and sampling rate via either re-segmentation of data or single-pass retraining of the HMMs.

4.2. Maximum-likelihood estimation of λ

By assuming that $G(F_0, \lambda)$ has a Gaussian distribution, we can find the maximum-likelihood estimate of λ straightforwardly. Its likelihood function is defined by the product of the Gaussian and the Jacobian of the transformation from F_0 to G. Maximizing the log likelihood function w.r.t. λ is equivalent to finding a λ that minimizes variance normalized by the geometric mean [8].

If the F_0 distribution has multiple modes, the Gaussian assumption is less appropriate. In such a case, Gaussianization [13] could be employed. Note that if the multiple modes of the F_0 distribution were caused by halving or doubling (i.e., F_0 extraction errors) we would not wish to model them!

5. EXPERIMENTS

We use a newly-recorded speech database of a semi-professional American female speaker having a standard accent uttered 1,130 CMU-ARCTIC sentences in a highly-controlled recording studio environment. The original sampling rate used for the recording was 44.1 kHz. The labels for the data were automatically generated from the word transcriptions and speech data using the Unisyn lexicon [14] and Festival's Multisyn voice building procedure [15]. The Multisyn procedure automatically identifies utterance-medial pauses, vowel reductions or reduced vowel forms as well as providing a phoneme segmentation.

5.1. HTS voice building

From the speech database and labels that include an initial phoneme segmentation, we train a set of speaker-dependent context-dependent multi-stream left-to-right MSD-HSMMs [16] that model three kinds

Table 1. Combinations used for speaker similarity evaluation. Analysis order M for each sampling rate was adjusted in advance to make the listening test compact.

Index	Type	f_s [kHz]	M	α
1	Recorded	44.1	n/a	n/a
2	Vocoded	44.1	69	ERB
3	Vocoded	22.05	64	ERB
4	Vocoded	16	49	Mel
5	HMM + Vocoded	44.1	69	ERB
6	HMM + Vocoded	22.05	64	ERB
7	HMM + Vocoded	22.05	64	Bark
8	HMM + Vocoded	22.05	64	Mel
9	HMM + Vocoded	16 + up-sampling	49	Mel
10	HMM + Vocoded	16	49	Mel
11	Recorded (different person)	16	n/a	n/a

of parameters for the STRAIGHT [17] mel-cepstral vocoder with mixed excitation (the mel-cepstrum, generalized $\log F_0$ and a set of band-limited aperiodicity measures) plus their velocity and acceleration features. An overview of the training stages is shown in Figure 2. First, monophone MSD-HSMMs are trained from the initial segmentation, converted to context-dependent MSD-HSMMs and re-estimated. Then, decision-tree-based context clustering is applied to the HSMMs and the model parameters of the HSMMs are thus tied. The clustered HSMMs are re-estimated again. The clustering processes are repeated until convergence of likelihood improvements and the whole process is further repeated using segmentation labels refined with the trained models in a bootstrap fashion.

5.2. Bootstrap voice building

In general, speech data sampled at higher rates requires higher order M for mel-cepstral analysis. We started by training models on lower sampling rate speech (16 kHz) with a low analysis order and gradually increased the analysis order and sampling rates via either re-segmentation of data or single-pass retraining of HMMs as shown in Figure 3. Minimum generation error (MGE) training [18] was employed in the final training stage.

5.3. Listening test

For assessing the effect of sampling rates and auditory scales on speaker similarity, we performed a listening test. The system configurations compared in the test are shown in Table 1. In addition to 44.1 kHz sampling rate natural speech, we evaluated vocoded speech and HMM-based synthetic speech for 16kHz, 22.05kHz, and 44.1kHz sampling rates. Up-sampling from 16kHz to 22.05kHz was also included. The mel, bark, and ERB scales were evaluated at the 22.05 kHz sampling rate. The analysis order M for each sampling rate was set using informal listening tests because including this factor in the formal listening test would have made the listening test too large. To remind listeners that their task was to judge speaker similarity, we included recorded speech from a different speaker (CLB from CMU-ARCTIC) having different characteristics from our target speaker. To evaluate similarity to the target speaker, 5-point comparison category rating (CCR) tests are used. The scale for the CCR test runs from 5 for "sounds like exactly the same person" to 1 for "sounds like a totally different person" and a few examples of recorded speech (44.1 kHz) from the target speaker are provided as a reference. English synthetic speech was generated for a set of 50

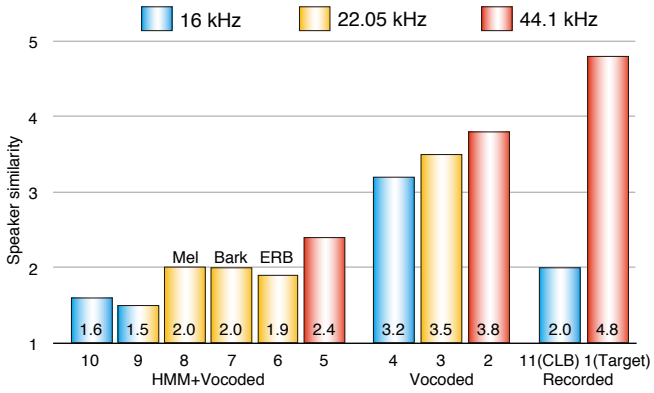

Fig. 4. Listening test results. The scale used is 5 for "sounds like exactly the same person" and 1 for "sounds like a totally different person." See index in Table 1 for details of each system.

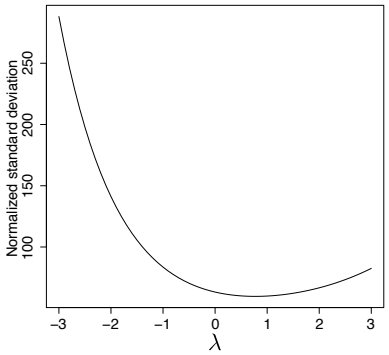

Fig. 5. The maximum-likelihood estimation of λ for the generalized logarithmic transform. All voiced frames in the database were used for the calculation.

test sentences randomly chosen from news and novel genres and the number of listeners was 23.

Figure 4 shows the result of the listening test. Since the listeners judged the target speaker's recorded speech as being from the same person (score 4.8) and CLB's recorded speech as being from a different person (score 2.0), we can be sure that they understood the speaker similarity task (and were not, for example, basing their judgements on naturalness). The baseline system (system 10) had a very low similarity score, although this system is an improved version of the 'HTS-benchmark system' used in the Blizzard Challenge, which has had good results every year. We summarize our conclusion regarding each factor as follows:

Sampling rate This is a very important factor. By downsampling from 44.1 kHz to 22.5 and 16 kHz, the scores drop from 3.8 to 3.5 and 3.2 for vocoded speech (system 2 to 4) and from 2.4 to 2.0 and 1.6 in HMM-based speech (system 5, 8, and 10). However by comparing systems 9 and 10 we can see that up-sampling of synthetic speech generated does not improve the similarity. In other words, the higher sampling rates enable better feature extraction.

Auditory scale This is a less significant factor. By comparing systems 6 to 8, we can see systems using the mel, bark and ERB scales have almost the same scores. Interestingly the highest score in HMM-based speech was obtained by system 5, even though that has the largest error in the auditory scale due to a high sampling rate and the use of the ERB scale.

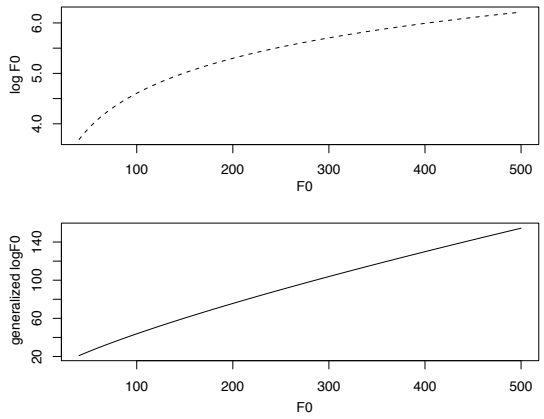

Fig. 6. Comparison of the logarithmic scale and the generalized logarithmic scale found using a maximum-likelihood estimate of λ.

5.4. Evaluation of generalized $\log F_0$

We performed a maximum-likelihood estimation of λ using all voiced frames in the database, before HMM training. Figure 5 shows the standard deviation normalized geometric mean for each value of λ (calculated at increments of 0.1), where the optimal value of λ is between 0.5 and 1.0. The optimal value obtained from the simplex method was 0.77. Figure 6 show the generalized logarithmic scale found using the maximum-likelihood estimator of λ, where we see the scale is more linear than a log scale.

Since synthetic speech using both the F_0 scales sounds natural and since only the difference is the dynamic properties of the F_0 trajectory, we considered that a MOS test of naturalness was not an appropriate way to evaluate the difference between the two scales. Instead we report the intra-utterance variance of the F_0 trajectory for each scale. We generated speech using the same set of test sentences as the previous listening test and compared the variance of F_0 values generated. Synthetic speech using the generalized logarithmic scale for F_0 had a 1.4 times larger intra-utterance variance in F_0 space than when using the normal logarithmic scale. This results in a more dynamic F_0 trajectory that one can easily perceive. A formal listening test evaluation of this result will be performed if a suitable methodology can be devised.

5.5. Audio examples

Audio examples for each method above are available online via `http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/meg.html`. We encourage interested readers to listen to these audio samples to hear the effect of these simple but effective methods.

6. CONCLUSIONS

In this paper we revisited some basic configuration choices made in HMM-based speech synthesis such as the sampling rate, auditory scale and logarithmic scale of F_0, which are typically based on experience from other fields. Contrary to what is generally accepted in ASR, higher sampling rates (above 16 kHz) lead to enhanced feature extraction and improved speaker similarity for speech synthesis. A generalized logarithmic transform of F_0 results in a wider intra-utterance variance of F_0 trajectories and more dynamic prosody.

Acknowledgements We thank James Coupe and his colleagues at the University of Washington, USA for use of the 'Megham' database. The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org). SK holds an EPSRC Advanced Research Fellowship. This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative.

7. REFERENCES

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[2] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, Sep. 2008.

[3] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia, Sep. 2008.

[4] E. Zwicker and B. Scharf, "A model of loudness summation," *Psych. Rev.*, vol. 72, pp. 2–26, 1965.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis," *IEICE Trans.*, vol. J87-D-II, no. 8, pp. 1565–1571, Aug. 2004, (in Japanese).

[6] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan (E)*, vol. 5, no. 4, pp. 233–242, Oct. 2000.

[7] J. Yamagishi *et al.*, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech 2009*, Brighton, U.K., Sep. 2009, pp. 420–423.

[8] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 2, pp. 211–252, 1964.

[9] K. Tokuda, T. Kobayashi, T. Fukada, H. Saito, and S. Imai, "Spectral estimation of speech based on mel-cepstral representation," *IEICE Trans. Fundamentals*, vol. J74-A, no. 8, pp. 1240–1248, Aug. 1991, in Japanese.

[10] K. Tokuda, T. Kobayashi, and S. Imai, "Recursive calculation of mel-cepstrum from LP coefficients," in *Technical Report of Nagoya Institute of Technology*, Apr. 1994.

[11] R. Patterson, "Auditory filter shapes derived with noise stimuli," *Journal of the Acoustical Society of America*, vol. 76, pp. 640–654, Mar. 1982.

[12] J. O. Smith III and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. on Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, Jul. 1999.

[13] S. S. Chen and R. A. Gopinath, "Gaussianization," in *NIPS 2000*, Nov. 2000, pp. 423–429.

[14] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech 1999*, vol. 2, Sep. 1999, pp. 823–826.

[15] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

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

[17] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "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, pp. 187–207, 1999.

[18] Y. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP 2006*, May 2006, pp. 89–92.