HMM-based sCost quality control for unit selection speech synthesis

Sathish Pammi[1], *Marcela Charfuelan*[2]

[1]ISIR, Universit Pierre et Marie CURIE (UPMC), Paris, France
[2]DFKI, Language Technology Lab, Berlin, Germany
sathish.pammi@isir.upmc.fr, marcela.charfuelan@dfki.de

Abstract

This paper describes the implementation of a unit selection text-to-speech system that incorporates a statistical model Cost (sCost), in addition to target and join costs, for controlling the selection of unit candidates. sCost, a quality control measure, is calculated off-line for each unit by comparing HMM based synthesis and recorded speech with their corresponding unit segment labels. Dynamic time warping (DTW) is used to perform such comparison at level of spectrum, pitch and voice strengths. The method has been tested on unit selection voices created using audio book data. Preliminary results indicate that the use of sCost based only on spectrum introduce more variety on style pronunciation but affects quality; whereas using sCost based on spectrum, pitch and voicing strengths improves significantly the quality, maintaining a more stable narrative style.
Index Terms: Text-to-speech synthesis, unit selection synthesis, statistical parametric synthesis, quality control

1. Introduction

Quality control of units in unit selection speech synthesis is a topic of high interest. Especially important are automatic approaches for finding the units that are intelligible and labelling error free for stable and good quality synthesis. Transcription and automatic labelling errors are the most frequent problems in unit selection synthesis. When we are dealing with large audio book corpora, the additional problem is handling the variable expressivity. The narrator in the audio book, might produce such a variability in speech style and pronunciation, that avoiding artifacts and abrupt changes in waveform concatenation is still a matter of research [1].

The use of HMM-based synthesis techniques to improve speech quality in unit selection is not a new topic. Several researchers have attempted to combine in a hybrid approach, statistical prediction of parameters with waveform concatenation. For example in [2, 3] a HMM-based unit selection approach is proposed, where acoustic parameters (spectral and fundamental frequency) generated with HMM models are used to guide the selection of units. This is done via sentence likelihood and a feature vector distance between HMM generated features and extracted features from the waveform unit candidates. A similar approach, using diphones as unit level, is adopted in [4], where a hybrid technique of unit selection from statistically predicted parameters is proposed. Also in [5] normalised distances between HMM trajectory and those of the waveform unit candidates are used for selecting final candidates in a unit sausage (lattice). The main difference in this last case, is an additional pruning strategy to generate a compact set of unit candidates.

In this paper a HMM-based synthesis approach is also used to improve unit selection speech quality. Like in the hybrid approach we use HMM-based trained models to generate acoustic parameters, but here we use those parameters off-line to precalculate a statistical model cost (sCost). Thus, the sCost is a measure of how different a sentence of the corpus is (in terms of acoustic parameters at level of units) from a sentence generated with statistically trained models (HMMs).

The sCost measure was developed in our previous work [6], where it was used to automatically find labelling errors, so to improve the quality of concatenation units. In this paper we extend our previous work in two ways: (i) sCost is used in addition to target and join costs for controlling the selection of unit candidates in a unit selection synthesiser; and (ii) sCost is calculated not only for spectral features but also for fundamental frequency and voicing strength features.

The objective is that the sCost model helps to discard units far beyond the average acoustics in the corpus and thereby contribute to select better quality units for concatenation. Additionally, since the HMM-based voice we use to generate parameters is trained with neutral style data, we expect that the sCost will penalise those segments (units) pronounced with a very different style. In some way, this approach is similar to the one described in [7], where synthetic speech data annotated as natural and unnatural is used to train a SVM model that helps to evaluate the naturalness of synthetic speech.

The paper is organised as follows. In Section 2 the methodology of sCost computation and its utilisation in unit selection synthesis is described. In Section 3 we describe how the neutral style HMM-based voice is created, the sCost model is calculated and how it is used in run time unit selection. In Section 4 the method presented in this paper is evaluated in a listening test, where a baseline unit selection voice is compared with two unit selection voices created with sCost model; main effects are discussed. Finally in Section 5, conclusions are made and future work is envisaged.

2. Methodology

The proposed methodology describes the usage of the HMM-based statistical model cost (i.e. sCost) in unit selection speech synthesis. We describe the procedure for estimating sCost from different parameters, using Dynamic Time Warping (DTW), and its use in selecting candidate units for synthesis.

2.1. Computation of sCost

As shown in Figure 1 the sCost is computed in several steps. As a first step, an automatic labeller estimates automatic segment labels based on recorded speech and phonetic transcription from text prompts. Secondly, an HMM voice is created by the HMM voice-building module using the automatic labels, generated in the previous step, and recorded waveforms. In the next steps, the HMM parameter generation module gen-

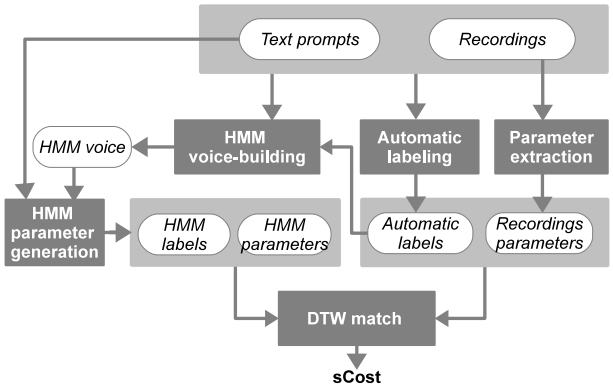

Figure 1: sCost computing methodology ([6])

erates parameters and HMM predicted segment labels from the text prompts. Having similar conditions for parameters dimension, frame size and frame-shift, parameters are extracted from the recorded waveforms. Finally, DTW computes an sCost by matching the extracted parameter feature vector sequence of the recorded speech and the generated parameters by the HMM parameter generation module. When aligning the two parameter vector sequences their corresponding unit segment labels are taken into account.

2.2. Unit selection using sCost

The unit selection based approach is based on: the selection of appropriate candidate units, which are close to the intended *target*, from a database of natural speech; and an appropriate combination of the selected units in order to achieve good speech quality. The unit selection algorithm plays a key role in identifying which of the available candidate units are appropriate for the target of intended speech to be synthesised.

According to the traditional unit selection algorithm [8], the algorithm includes two types of costs: *target cost* to define how well a candidate unit from the database matches the target unit; and *concatenation cost* to define how well two selected units can combine at joints. The cost functions can be written as the following:

$$targetCost(u_i) =< w, c(u_i) > \quad (1)$$

$$joinCost(u_i, u_{i-1}) =< w, c(u_i, u_{i-1}) > \quad (2)$$

where u_i is the candidate unit i; c is the cost vector containing several feature costs; and w is the weight vector for the features.

In the proposed method, each candidate is associated with a precomputed sCost (i.e. quality measure) for each parameter. The parameter specific sCost measures can be combined as the following:

$$sCost(u_i) = W^{\mathrm{T}} * \begin{pmatrix} sCostPAR_1(u_i) \\ ... \\ sCostPAR_n(u_i) \end{pmatrix} \quad (3)$$

where W is a weight vector; sCostPAR represents a parameter specific sCost measure.

The overall cost for selecting units in the dynamic programming stage can be modified as the following:

$$totalCost(u_i) = W_1^{\mathrm{T}} * \begin{pmatrix} targetCost(u_i) \\ joinCost(u_i, u_{i-1}) \\ sCost(u_i) \end{pmatrix} \quad (4)$$

At the stage of selecting units, the dynamic programming algorithm finds the best suitable candidates for the target by minimising the total cost function described above. Beam search is used to minimise the speed of computation.

3. Realisation

In order to test the method proposed in this paper several unit selection voices were created using the MARY TTS voice building tools [9]. One HMM-based voice and three unit selection voices were created using audio book data, in this case "Mansfield Park" released in the Blizzard Challenge 2013 [1]. The audio book data was already split into prosodic phrase level chunks. The sentence segmentation and orthographic text alignment of the audio book has been performed using an automatic sentence alignment method – LightlySupervised – as described in [10].

3.1. HMM-based voice building

HMM-based voices are well known to produce flat spectral trajectories and smooth F0 contours, which for our purposes will be a good approximation of the context-dependent average segment acoustics. Additionally, and in order to generate a HMM-based voice with a stable, not so expressive narrative style, we have used the same techniques used in [11] to create a neutral voice out of audio book data. That is, we have extracted acoustic features from each sentence of the corpus and perform principal component analysis so to discard sentences beyond a PC1 threshold. For this experiment we have extracted the following acoustic features:

- Fundamental frequency (F0) and F0 statistics: mean, max., min., and range.
- Number of words.
- Average energy, calculated as the short term energy averaged by the duration of the sentence in seconds.
- Voicing rate calculated as the number of voiced frames per time unit.
- Five band pass voicing strengths estimated with peak normalised cross correlation of the input signal.

For calculating voicing strengths, the input signal is filtered into five frequency bands and mean statistics of these measures are extracted per sentence. Voicing strengths features are normally extracted in the MARY TTS voice building framework for HMM-based synthesis using mixed excitation.

As in [11], we have found that also in this data, voicing rate and voicing strengths contribute more than F0 or MFC to the variance of the first principal component. This might indicate that the data contains more variation in speaking styles (voice quality) than extreme emotions. Using this method we have selected 3363 sentences out of the aprox. 7000 sentences of the whole audio book corpus, for building a HMM-based neutral voice. When creating a HMM-based voice in the MARY TTS framework, three types of acoustic features are extracted:

- MFC: Mel generalised cepstrum, dimension 25, extracted using SPTK [12],

- LF0: Log fundamental frequency, dimension 1, extracted using snack [13],
- STR: Voicing strengths, dimension 5, from 5 bands of frequency, extracted using snack and a set of filters provided in the MARY TTS voice building framework.

For the experiments with sCost model, these features were also extracted from the whole corpus, with which we created three unit selection voices, two of them employing sCost, as explained below.

3.2. sCost estimation for audio book data

DTW, a dynamic programing technique with optimal alignment to match the acoustically most similar sections between two phonetic segments, is implemented in MARY TTS for estimation of sCost between extracted parameters from recordings and generated parameters from the HMM voice. Here, an automatically labelled phone segment in the recorded speech is matched with the corresponding segment generated by the HMMs. The criterion for finding the optimal path is the Mahalanobis distance between the recorded and generated parameter vectors (i.e. MFC, STR, LF0), using the variance computed per phone on the recorded waveforms. sCost is computed as the sum of the Mahalanobis distance over the optimal path, divided by the number of frames in the recorded segment and in the generated segment.

MARY TTS unit selection uses diphones as basic units. An average of two half-phone sCost measures are considered as the diphone's sCost. In order to estimate sCost for each half-phone, the acoustic parameters are also extracted from the whole corpus of aprox. 7000 sentences, and generated using the HMM parameter generation component of the neutral HMM-based voice. In this work, we compute three sCost measures for each unit. They are: sCostMFC using MFC parameters; sCostSTR using STR parameters; sCostLF0 using LF0 parameters.

3.3. Unit selection voice building

The unit selection voice building use the standard approach in MARY TTS framework [14]. The only difference in the new voices is that they contain precomputed sCost measures in timeline files. All the precomputed measures are put into a timeline file, together with other timeline files in the unit selection voice.

As mentioned before, for testing the method presented in this paper we have created three unit selection voices, using the whole corpus (aprox. 7000 sentences), with the following characteristics:

- voice A: baseline voice, it does not use sCost model,
- voice B: a unit selection voice that uses a sCost model calculated with only MFC features, as in [6],
- voice C: a unit selection voice that uses a sCost model calculated with MFC, STR and LF0 features.

For run time synthesis, the MARY TTS unit selection algorithm combines the usual steps of pre-selecting candidate units, a dynamic programming phase combining weighted join costs and target costs, and a concatenation phase joining the selected units into an output audio stream. In the current version, a very small pre-selection tree is manually specified and can pre-select units, e.g., by their phone or diphone identity [14]. A beam search is used in the dynamic programming step to keep processing time low.

In addition to join costs and target costs, we add statistical model costs to the phase of dynamic programming phase as described in Eq. 4. Total sCost in this equation is measured with weighted sum of parameter-specific sCosts as the following:

$$sCost(u_i) = W_0^{\mathrm{T}} * \begin{pmatrix} sCostMFC(u_i) \\ sCostSTR(u_i) \\ sCostLF0(u_i) \end{pmatrix} \quad (5)$$

The weights of join cost, target cost and statistical model cost are tuned manually, "heuristically", for each voice based on subjective perception. For Voice A (no sCost), the weights for sCost becomes zero. For Voice B (sCost MFCs), the weights of sCostSTR and sCostLF0 becomes zero. To make a fair comparison, we manually tuned weights of all three voices to their best performance by listening to the synthetic speech of several random sentences.

4. Evaluation

Since audio book data is more expressive, it is difficult to define an objective measure, like spectral distance, to compare sentences that can be correctly pronounced in different ways. So in order to evaluate the effect of sCost we have performed a preference perceptual test, where we ask users to listen and compare pairs of sentences and select the one that in their opinion sounds better in quality and pronunciation for the given text. As test sentences we have selected 12 sentences from another book: "The adventures of Tom Sawyer", for reference we have included them in Figure 1.

As an example of the effect of sCost on the generated sentences, we can see in Figure 2 the F0 contour obtained with the three unit selection voices A, B and C for sentence 6. `Tom, what on earth ails that cat?`. We can observe that in this example the F0 contour generated with voice C is much more smooth than the contours generated with voices A and B. Perceptually, the sentence generated with voices A and B present much more variations in pronunciation, but with introduction of artifacts that degrade the quality. The sentence generated with voice C, on the other hand present a more stable narrative style, with better quality.

A more detailed, spectral view of the word `ails` in the same sentence, is presented in Figure 3. In this figure we can observe how the sCost model in the sentence generated with voice C, present a considerable reduction in spectral discontinuities. These observations seem to correlate with the results obtained in the listening test.

4.1. Listening test

Seventeen people participated in the listening test, among them several speech experts and most of them non-native speakers of English. The users listened in random order 12 pairs of sentences, in three sessions: AB, AC and BC. Where AB means that users listened the 12 sentences generated by unit selection synthesisers A and B.

As shown in Figure 4, the results indicate that broadly, subjects preferences are:

- voice A (72%) over voice B (28%),
- voice C (78%) over voice B (22%),
- voice C (58%) over voice A (42%).

Although overall preferences for voice A and C are high, subjects clearly indicate their preference towards 8 samples of

1. Well I WILL, if you fool with me.
2. Tom knew that when his name was pronounced in full, it meant trouble.
3. Huckleberry Finn was there, with his dead cat.
4. It was on a hill, about a mile and a half from the village.
5. The boys clasped each other suddenly, in an agony of fright.
6. Tom, what on earth ails that cat?
7. Some people think they're mighty smart, -- always showing off!
8. They had a famous fried-egg feast that night, and another on Friday morning.
9. I want to go home.
10. The stillness continued; the master searched face after face for signs of guilt.
11. Becky's face paled, but she thought she could.
12. The village was illuminated; nobody went to bed again; it was the greatest night the little town had ever seen.

Table 1: Test sentences used in in the listening test.

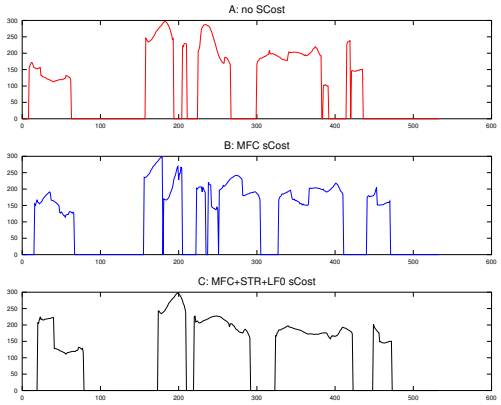

Figure 2: F0 contours of test sentence 6. "Tom, what on earth ails that cat?" generated with unit selection voices A, B and C.

voice C, 3 samples of voice A, and one sample for which they have almost equal preference.

4.2. Discussion

Among the synthesised examples[1], sentence 6 with text `Tom, what on earth ails that cat?` also reveals some interesting insights of the approach. The synthesised audio from system A is realised as `Tom, what earth ails that cat?` (deletion of "on"), whereas it is realised by system B as `Tom, what but on earth ails that cat?` (insertion of "but"). These errors are mostly due to misalignment in automatic labelling. However, such problems were successfully avoided by the realisation in system C. This means that sCost computed with all parameters, seems to deal with automatic labelling errors appropriately.

The average consecutive length (ACL) of each unit selection system are:

$$\begin{aligned} ACL_A &= 6.2 \\ ACL_B &= 3.1 \\ ACL_C &= 5.0 \end{aligned}$$

while the average consecutive length of units in system A is much higher than in systems B and C, it is much lower for system B. This means that the insertion of sCost into the unit selection algorithm reduces the average number of units that are consecutive, specially when sCost is precomputed for MFC only. In the listening test, reduction of ACL have had a negative effect on the performance of system B, because more dissimilar joins introduced more perceptible artifacts, that reduced the speech quality.

When the sCost uses all parameters including MFC, STR and LF0, the average consecutive length in unit selection is increased. Interestingly, the subjective preference is higher for system C when compared to system A, though the average consecutive length is lower. Therefore, it seems that system C is maintaining a fair balance between the consecutive selection of units and acoustically similar units.

A counter example is the following: in the synthesised sentence 9, `I want to go home`, the subjects fully preferred system A over system B and system C over system B. However, 75% of subjects preferred system A instead of system C. The average consecutive length of this particular sentence synthesised by systems A, B and C are 8.67, 2.89 and 5.2 respectively. The choice of voice A in this particular case might be due to less number of joins in the synthesised audio. Thus, we can conclude that, although sCost helps to reduce concatenation errors and make the voice style more stable, these type of errors still appear, so the approach will be further investigated in order to improve the join model in combination with sCost.

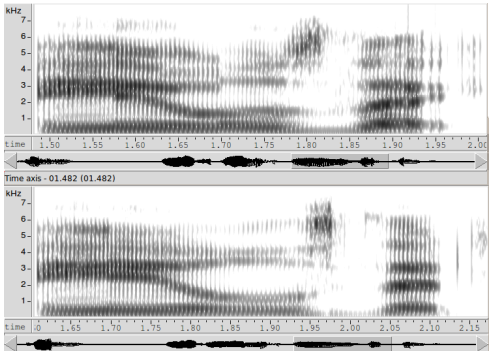

Figure 3: Spectrum of a section of sentence 6. corresponding to the word "ails" generated with unit selection voices A (upper spectrum) and C (lower spectrum).

5. Conclusions

In this paper we have presented the implementation and evaluation of a unit selection text-to-speech system that incorporates

[1] http://www.dfki.de/~charfuel/listening_test/listening_test.html

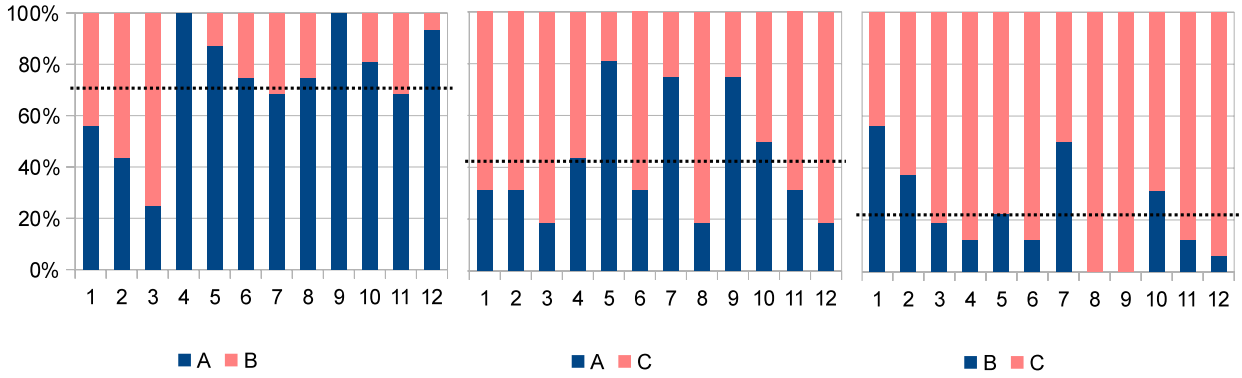

Figure 4: Listening test results: preference for 12 sentences synthesised with systems A and B, A and C and B and C. Dashed line in figures indicate the average preference between systems.

a statistical model cost (developed in a previous work [6]), in addition to target and join costs, for controlling the selection of unit candidates.

We have extend our previous work in two ways: (i) sCost is used in addition to target and join costs for controlling the selection of unit candidates in a unit selection synthesiser; and (ii) sCost is calculated not only for spectral features but also for fundamental frequency and voicing strength features. The method has been tested on unit selection voices created using audio book data. Due to the highly variable expressivity of the data, the HMM-based voice used to calculate sCost was built with neutral style data, automatically selected from the corpus.

Three unit selection voices were created, using all the data in the audio book, to perform a listening test where a baseline system without sCost was compared against: a system using a MFCsCost; and another using MFCsCost, STRsCost and LF0sCost. The listening test results indicate a clear preference for the system that include the three types of sCost. We have also discussed and presented examples of the effect of sCost on the F0 contour and spectrum, as well as, the effect on the average consecutive length of units.

We have shown how the use of sCost based only on spectrum introduce more variety on style pronunciation but affects quality; whereas using sCost based on spectrum, pitch and voicing strengths improves significantly the quality, maintaining a more stable narrative style. In future work we will not only investigate a better join model that suits for this approach, but also work towards a generic approach for style control using the proposed statistical model cost measures.

6. Acknowledgements

This work is supported by the European Union Seventh Framework Programme under grant agreement n288241 through the Michelangelo project. This work is also supported by the EU project SSPNet (FP7/2007-2013) and partially supported by AVATAR 1.1 project.

7. References

[1] S. King and V. Karaiskos, "Blizzard Challenge 2013," http://www.synsig.org/index.php/Blizzard_Challenge_2013.

[2] Z.-H. Ling and R.-H. Wang, "HMM-based hierarchical unit selection combining kullback-leibler divergence with likelihood criterion," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, 2007, pp. IV–1245–IV–1248.

[3] Z.-H. Ling, L. Qin, H. Lu, Y. Gao, L.-R. Dai, R.-H. Wang, Y. Jiang, Z.-W. Zhao, J.-H. Yang, J. Chen, and G.-P. Hu, "The USTC and iFlytek speech synthesis systems for blizzard challenge 2007," in *Proceedings of Blizzard Challenge 2007*, Bonn, Germany, 2007.

[4] A. W. Black, C. L. Bennett, B. C. Blanchard, J. Kominek, B. Langner, K. Prahallad, and A. Toth, "Cmu blizzard 2007: A hybrid acoustic unit selection system from statistically predicted parameters," in *Proceedings of Blizzard Challenge 2007*, Bonn, Germany, 2007.

[5] 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," in *Proceedings of Blizzard Challenge 2010*, Kansai Science City, Japan, 2010.

[6] S. Pammi, M. Charfuelan, and M. Schroder, "Quality control of automatic labelling using HMM-based synthesis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4277–4280.

[7] H. Lu, Z.-H. Ling, L.-R. Dai, and R.-H. Wang, "Building HMM based unit-selection speech synthesis system using synthetic speech naturalness evaluation score," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 5352–5355.

[8] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 373–376.

[9] MARY TTS, "VoiceImportTools Tutorial," https://github.com/marytts/marytts/wiki/VoiceImportToolsTutorial, 2012.

[10] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Interspeech*, Makuhari, Chiba, Japan, 2010.

[11] M. Charfuelan and I. Steiner, "Expressive speech synthesis in MARY TTS using audiobook data and EmotionML," in *Proceedings of Interspeech 2013*, Lyon, France, 2013.

[12] S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, and H. Zen, "Speech signal processing toolkit (SPTK), Version 3.3," http://sp-tk.sourceforge.net, 2009.

[13] K. Sjölander, "The snack sound toolkit," http://www.speech.kth.se/snack, 2012.

[14] M. Schröder, S. Pammi, and O. Türk, "Multilingual mary tts participation in the blizzard challenge 2009," in *Proc. Blizzard Challenge*, vol. 9, 2009.