The GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach

Antti Suni[1], *Tuomo Raitio*[2], *Martti Vainio*[1], *Paavo Alku*[2]

[1]Department of Behavioural Sciences, University of Helsinki, Helsinki, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
antti.suni@helsinki.fi, tuomo.raitio@aalto.fi

Abstract

This paper describes the GlottHMM speech synthesis system for Blizzard Challenge 2012. The aim of the GlottHMM system is to combine high-quality vocoding and detailed prosody modeling in order to produce expressive, high quality, synthetic speech. GlottHMM is based on statistical parametric speech synthesis, but it uses a glottal flow pulse library for generating the excitation signal. Thus, it can be regarded as a hybrid system using the pulses as concatenative units that are selected according to the statistically generated voice source feature trajectories. This year's speech material was challenging especially, but despite that we were able to achieve a clean, intelligible voice with decent above average prosody characteristics.

Index Terms: statistical parametric speech synthesis, hybrid, unit selection, glottal inverse filtering, glottal flow pulse library

1. Introduction

The Blizzard challenge 2012 was a definite step up from the previous ones, involving under-researched topics such as suboptimal recordings and recording conditions, continuous speech, prose with mixed styles and synthesizing paragraph-length utterances. Within limited time, none of the advanced topics could be given proper attention on our submission, and even achieving acceptable level of intelligibility was not a trivial task this year. Nevertheless, we found it beneficial to participate in the challenge in order to test new ideas and explore the limits of our system, designed and evaluated previously on studio quality and rather formal speech.

The paper is organized as follows. Section 2 describes the aim of our research and gives an overview of the system. Section 3 describes the methods used in front-end and voice building. Section 4 describes feature extraction, parameter training and generation, and synthesis. The results of the evaluation are described in Section 5 and Section 6 summarizes the findings.

2. Overview of the system

The overall aim of the GlottHMM research is to combine novel vocoding methods and detailed prosody modeling in order to produce expressive, high quality, synthetic speech. The overview of the text-to-speech (TTS) system is shown in Figure 1. In prosody modeling, our general methodology is strong coupling with linguistic front-end and hidden Markov model (HMM) training; iterative refinement of HMM and contextual labels. Central to our prosody modeling is the concept of word prominence, annotated automatically for training corpus, and used as a contextual feature in HMM training. In order to model expressive prosody, especially on paragraph sized utterances, good predictive features are needed. In addition to part-of-speech, we typically use such linguistic features as (noun) phrase structure, focus particles, word order and discourse information status, as well as numerical features derived from automatic annotation process and text data. These features, with fairly indirect relationship with acoustic parameters, are used only on predicting the symbolic prosody labels like prominence and breaks, not as contextual features in HMM training.

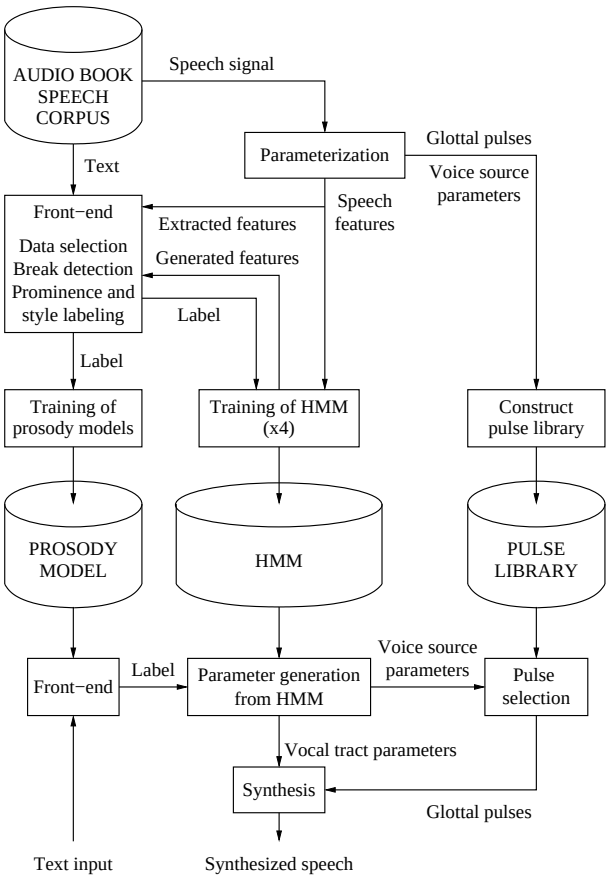

Figure 1: *Overview of the TTS system.*

The system uses a vocoder [1, 2, 3] that combines methods both from statistical parametric and unit-selection systems. The goal of this approach is to maintain the flexibility of statistical synthesizers while reaching the high quality of speech of the unit-selection systems. Contrary to usual hybrid systems, our synthesizer does not need a huge speech database or a huge amount of stored speech units (e.g. diphones), but it requires only a relatively small number of natural glottal flow pulses extracted from a small part of the speech database.

The vocoder first parametrizes speech into vocal tract and voice source features using glottal inverse filtering. The purpose of this decomposition is to accurately model both of the speech production components: source and filter. The voice source is converted into a glottal flow pulse library to enable reconstruction of the natural voice source in the synthesis stage. The pulse library consists a small number of (automatically) selected glottal flow pulses linked with the corresponding glottal source features. In synthesis stage, the voice source is reconstructed by selecting glottal flow pulses from the pulse library according to the voice source features. Thus, the voice source preserves the characteristics of natural excitation, such as spectral tilt, phase spectrum, and the fine structure of the excitation noise. Once the voice source is created, it is filtered with the vocal tract filter to generate speech.

This type of system is very flexible; the statistically modeled parameters can be easily adapted as in pure statistical TTS, and the small pulse library can be easily changed or modified according to, for example, speaker or speaking style. The cost of such a system is the addition of several voice source related parameters that need to be trained and the increased time in creating the excitation signal.

3. Voice building

In this section, we will discuss the steps of our voice building process, second time applied to English. Compared to e.g. Festival, there is a closer relationship with label generation and the training speech in our process, involving steps of partial HMM training and use of acoustic features in labeling.

3.1. Data selection and linguistic features

Three audio books by Mark Twain read by a male amateur reader were given as training material for the challenge. We decided to use just one book, assuming that consistent style and recording conditions would compensate for the smaller size of the training data. “Adventures of Tom Sawyer” was selected, on the grounds that it contained the least amount of out-of-vocabulary (OOV) words and foreign names with possibly unpredictable pronunciation. In retrospect, this choice was a grave mistake as later listening revealed the other two books to be much better in terms of recording quality and background noise. Of the selected book, we chose only the utterances with provided confidence scores of 100 percent, and we also skipped the sentences containing OOV words. The final size of the pruned training data was 3740 sentences.

External tools were used for initial labeling. Pronunciations and syllabification was performed with Unilex lexicon, general American variant. For part-of-speech (PoS) labeling and syntactic chunking, TreeTagger [4] was applied. PoS tags were used to disambiguate pronunciations.

3.2. Phrase breaks

Phrase breaks were acquired from original speech data. Firstly, monophone HMMs were trained with silence models attached to punctuation symbols. Secondly, the utterances were aligned with optional silences after each word. The recognized silences were further divided into three categories of phrase boundary strength. The categories were determined using silence duration and the duration of final syllable before silence. In synthesis, the breaks were predicted by rule, mainly based on punctuation.

3.3. Phrase style

While listening to the audio book, we identified three general reading styles suitable for modeling:

1. Low pitched, rather monotone, often suspenseful passages
2. Normal narrative style with rather lively prosody
3. Often high pitched, lively quotations

Quotations themselves were found to be very heterogeneous, with reader acting various characters, but more finer grained classification seemed out of reach, especially for prediction purposes. The styles were annotated by first training a voice without style labels, an “average-style” voice. Then parameters, fundamental frequency (F0) and energy of the training utterances were generated and compared to the original ones, with the idea that energy and F0 would on average be higher in the original utterances for quotations and lower for suspenseful style. The raw style score for each utterance was calculated as the weighted sum of differences between original and generated mean values of F0, energy and harmonic-to-noise ratio (HNR). The raw values were further binned into three classes, corresponding to the aforementioned styles. The weights and division points were set by hand after some experimentation.

Alternatively, we considered just skipping the utterances deviating considerably from the generated parameter trajectories. This would have probably resulted in a more stable voice but with no options to model styles in synthesis time.

Unfortunately, we neglected the work on various typographical conventions on marking quotations in text. Thus, in synthesizing test utterances, we were not able to predict but few phrases to be uttered in quotation style.

3.4. Word prominence

Word prominence was determined using similar approach as in annotating the utterance style, first training a voice with simple set of contextual features and then comparing original (O) and generated (G) acoustic-prosodic parameters [5]. Compared to Blizzard Challenge 2010 [2], we are moving towards simpler, less supervised method, requiring only setting of weights of parameter types, but no manual labeling. The proper set of parameters and measurements is still under development, but we know that for example F0 in Finnish correlates with perceptual prominence so that the higher the peak and the faster and larger the movement, the more prominent the word is perceived [12]. For the current entry, mean and variance normalized measures were made for F0, energy, HNR and duration. To detect local syllable peaks, we calculated the difference between previous and current syllable mean (rise) and difference between current and next syllable mean (fall), as well as the mean value of the current syllable, normalized over a window of five syllables (max). These were calculated for both O and G parameters, and the differences between O and G of rise, fall and max were also calculated. Mean values were used instead of minimums and maximums because they are more robust to e.g. octave jumps. After some experimenting the weights of parameter types were set manually as F0 = 0.5, energy = 0.25, duration = 0.25. HNR did not seem to contribute probably due to noisiness of the data. Both O and diff(O,G) measurements were taken into account with equal weights. Rise, fall and max were also given equal weight. The sum of these normalized measurements (see Fig. 2) was the calculated and binned to four classes, corresponding roughly to unaccented, secondary accent, primary accent and

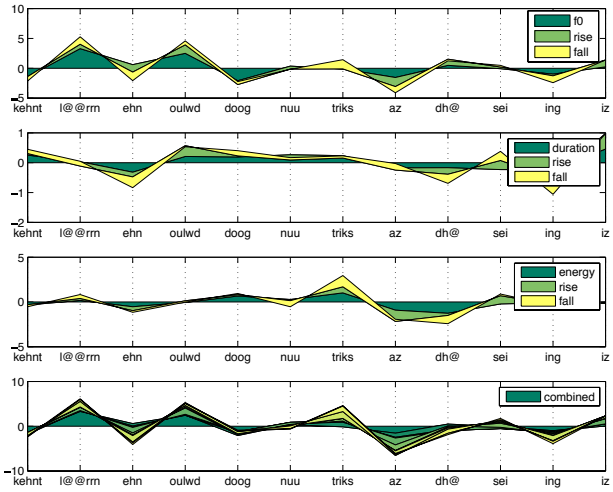

Figure 2: *Example of prominence annotation of complex sentence with two contrasts:"Can't learn an old dog new tricks, as the saying is"*

emphasis. Only lexically stressed syllables were taken into account for word prominence labels.

The general experience was that our prominence annotation method did not work very well here compared to previous experiments with more formal speech. The larger F0 movements, especially on quotations, seemed often to be more related to higher level discourse factors than prominence signaling.

In generating labels for the test utterances, word prominence was predicted by a combination of classification and regression tree (CART) and rules. The CART was trained on automatically annotated training data, the same data as used for voice building. The features used for training the model were average prominence of the word base form in the training data, part-of-speech, and information content with window length five. For words with few instances in the training data (< 5), the average prominence of the part-of-speech class was used instead of average prominence of the word. The model included also phrase style, as well as features describing the position of the word in phrase and utterance.

Additional rules were included in an attempt to handle some rare phenomena that could not be learned from relatively small, noisy training data. These included discourse related factors for synthesizing contextually appropriate prosody in paragraph-sized chunks:

- Decrease prominence of the previously seen (given) noun if it is part of complex noun phrase
- Increase prominence of potentially contrastive adjective modifiers if the head is given
- Increase prominence of the first content word of the paragraph
- Increase prominence of words with all-capital spelling
- Disallow many high-prominence words after the main verb, save last

4. Training and Synthesis

4.1. Feature extraction

The parametrization of the GlottHMM vocoder is illustrated in Figure 3. The speech signal s(n) is first high-pass filtered in

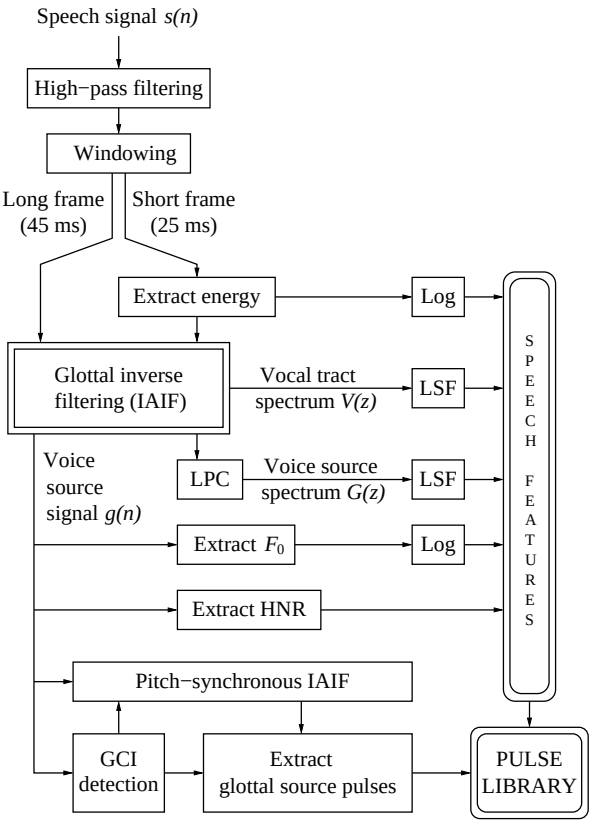

Figure 3: *Illustration of the parametrization stage.*

order to remove possible low-frequency fluctuations, and then windowed into two types of frames. A short frame (25 ms) is used for measuring the energy of the speech signal, after which glottal inverse filtering is applied in order to estimate the vocal tract filter V(z) and the voice source. The estimated voice source is parameterized with spectral tilt G(z) measured with an all-pole filter. A modified version of the iterative adaptive inverse filtering (IAIF) [6, 7] is used for estimating the vocal tract filter and the voice source. Linear predictive coding (LPC) is used for estimating the spectra inside the method. Both the vocal tract filter V(z) and spectral tilt of the voice source G(z) are converted to line spectral frequencies (LSF) for enabling robust statistical modeling.

The longer frame (45 ms) is used for extracting the other voice source features which require that the frame includes multiple fundamental periods even with very low F0. The estimated voice source signal g(n) is used for defining the F0, estimated with an autocorrelation method. The the harmonic-to-noise ratio (HNR) indicates the degree of voicing in the excitation, i.e., the relative amplitudes of the periodic vibratory glottal excitation and the aperiodic noise component of the excitation. The HNR is based on the ratio between the upper and lower smoothed spectral envelopes (defined by the harmonic peaks and interharmonic valleys, respectively) and averaged across five frequency bands according to the equivalent rectangular bandwidth (ERB) scale. The speech features extracted by the vocoder are depicted in Table 1.

The glottal closure instants (GCI) of the voice source g(n) are estimated with a simple peak picking algorithm that searches for the negative excitation peaks of the glottal flow

derivative at fundamental period intervals. Only the peaks that are approximately at fundamental period away from each other are accepted as GCIs. For all the found two-period speech segments, the modified IAIF algorithm is applied pitch-synchronously again in order to yield a better estimate of the glottal flow. The re-estimated glottal flow pulses are windowed with the Hann window, and a glottal flow pulse library is constructed from the extracted pulses and the corresponding voice source parameters.

4.2. Pulse library

The construction of the pulse library is performed separately from the training. Only 1–10 short speech files is enough to estimate enough glottal flow pulses to the library, consisting usually from 1000 to 20 000 pulses. The size of the pulse library can be greatly reduced for example by k-means clustering and selecting the centroid pulses or by including only the most commonly used pulses, estimated by synthesizing several speech files and counting the usage of the pulses. Moreover, different pulse libraries can be used for synthesizing different voices or voice qualities.

For the present voice, the pulse library was built from ten diverse utterances selected for phonetic and F0 range coverage. The pulse library contained a total of 22 414 pulses. The size of the pulse library was not reduced since the synthesis time was not an issue here. The individual weights for the voice source features for selecting the pulses were set to 0.5, 0.2, 1.0, 0.2, and 1.0 for vocal tract spectrum, glottal flow spectrum, HNR, energy, and F0, respectively. Vocal tract spectrum was included in the weights since it is also a good cue for certain voice types. Target and concatenation cost weights were set to 1.0 and 2.0, respectively. All the weights were mostly tuned by hand.

4.3. Parameter training

After the annotation steps, contextual features including word prominence and phrase style labels were extracted, and HMMs were trained in a standard HTS fashion [10], except that five iterations of MGE-training [11] was included for vocal tract LSFs as a final step. LSF and energy features were trained together in a single stream in order provide better synchrony between the parameters. Other features were trained in individual streams except F0 which uses a multi-space distribution (MSD) stream.

First experiments provided fairly unstable and muffled synthesis quality, indicating alignment problems in training. Since LSFs are highly correlated with each other, there are known problems in training them. To remedy the situation, we opted to use the differential of the LSFs [8] for vocal tract parameterization. The LSF training vector thus contained 31 values, of which the first one was the first LSF, next 29 were the differences between the adjacent LSFs, and the last one was the distance of the last LSF to π. In order to get the distributions of the differential LSFs more Gaussian, the square root of the distances were used for training. In parameter generation, the differential LSFs were equalized so that the sum of the 30 first LSF distances matched the 31th, the distance to π.

Table 1: Speech features and the number of parameters.

Feature	Parameters
Fundamental frequency	1
Energy	1
Harmonic-to-noise ratio	5
Voice source spectrum	10
Vocal tract spectrum	30–50
Pulse library	10–20 000 pulses

4.4. Parameter generation

Examining the trees, questions concerning the phrase style appeared quite early, causing fragmentation. Low suspenseful style sounded good, but the normal narrative style was too enthusiastic and jumpy. To stabilize the voice, we considered adaptive training approach, but settled for just combining low and normal styles because of tight schedule. With these changes, we obtained fairly intelligible final voice, yet still in need of lot of post-filtering (formant enhancement) to reduce the averaging effect. The test sentences were synthesized applying both parameter generation considering global variance (GV) and post-filtering. Looking at the results, the post-processing probably went too far and some internal listening would have been in order. There was also a harsh, high-frequency noise present in our voice. This was present already in original recordings, but became more distracting in heavily processed synthesis. Noise reduction should perhaps been applied to training utterances. The differential LSFs with GV seemed to also contribute to the problem, finding non-existing formants in high frequency regions. GV was then selectively applied to only lower order LSF coefficients, but the harsh quality still remained. In order to save listener's ears, some room reverberation was added to final synthesized paragraphs, hoping it would smooth the voice quality, but the results indicate that this had not much effect.

4.5. Synthesis of speech waveform

The flow chart of synthesis stage is shown in Figure 4. In synthesis, the voice source is reconstructed by selecting and concatenating pulses from the pulse library that yield the lowest target and concatenation costs given the voice source parameters. This process is optimized with Viterbi search for each continuous voiced segment.

Minimizing the target cost ensures that a pulse with desired voice source characteristic, such as fundamental period, spectral tilt, and amount of noise, is most likely to be chosen. The target cost is the error between the voice source features generated from HMMs and the ones that are linked to pulses in the pulse library. The target cost is composed of the mean square error of each feature, normalized by mean and variance across the pulse library, and weighted by individual target cost weights for each feature. Minimizing the concatenation error ensures that adjacent pulse waveforms are not too different from each other, providing a smooth speech quality without abrupt changes. The concatenation error is the mean square error between adjacent pulse waveforms. In order to prevent selecting the same pulse in a row, leading to buzzy excitation, an small bias is introduced to the concatenation cost of the pulse with itself.

The target and concatenation costs can be weighted individually to produce a smooth but accurate excitation. After the selection, the pulses are scaled in energy and overlap-added according to fundamental frequency to create a continuous, natural-like excitation. Although the selection process will most likely select pulses with approximately correct fundamental period, the pulses can be optionally interpolated to correct length. An example of the excitation and the resulting speech signal is shown in Figure 5.

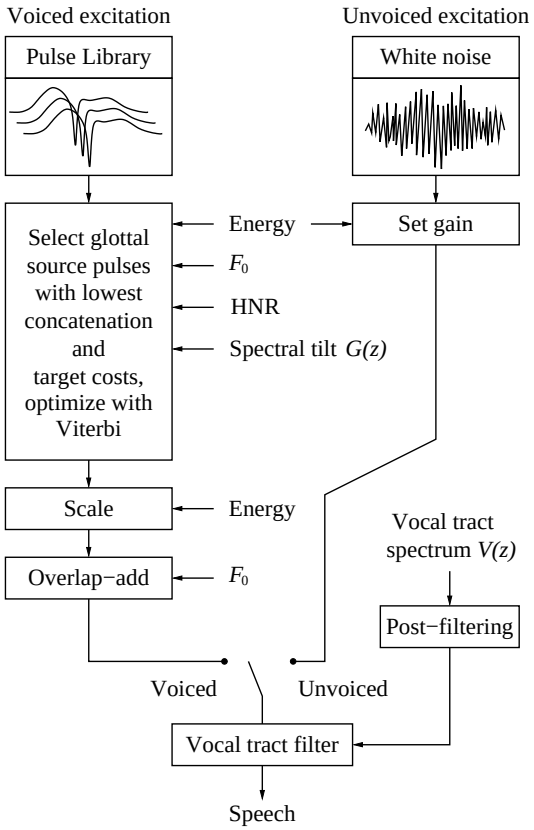

Figure 4: *Illustration of the synthesis stage.*

The unvoiced excitation is composed of white whose gain is determined according to the energy measure from HMMs. Formant enhancement [9] is applied to the vocal tract LSFs in order to alleviate for the over-smoothing of statistical modeling. Finally, LSFs are converted back to LPC coefficients describing the vocal tract spectrum V(z) and used for filtering the combined excitation.

5. Results and discussion

5.1. MOS, similarity and intelligibility

As expected, the results on naturalness and similarity of our submission were low, being on the lower third group of all submissions. The naturalness of our system was hurt by the wrong choice of training material and the overcompensation of the initial bad quality by post-processing, resulting in artificial tone of voice. The similarity score was additionally affected by the selection of original utterances in listening test, which were similar in recording and speech quality to the two books excluded in training our system.

With the help of more direct modeling of formants with differential LSFs, we were able to achieve top intelligibility, but then again, the other scores were probably adversely affected by losing the exact positions of LSFs. The intelligibility results are shown in Fig. 6. Our system is marked with letter D.

5.2. Paragraphs

The interesting part of this year's challenge was the synthesis and fine grained listening test of audio book paragraphs.

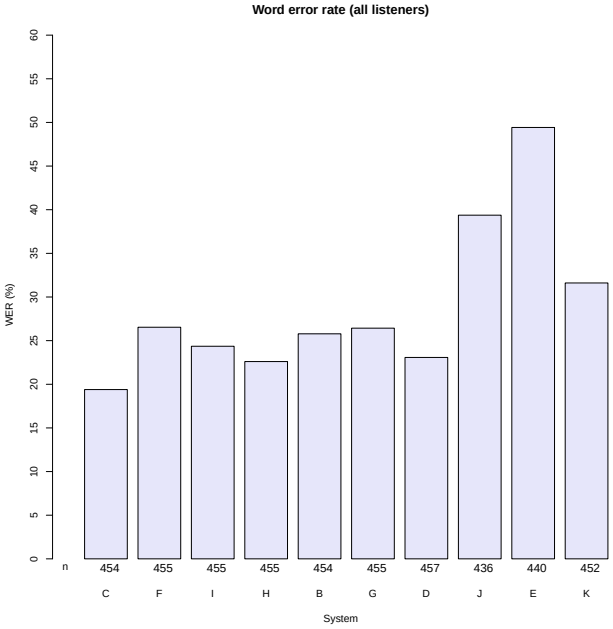

Figure 6: *Intelligibility results (D - GlottHMM).*

The questions asked from listeners covered specific aspect of prosody as well as overall quality. Here, apart from pleasantness, which was scored low, our system fared better. Especially, our prominence and break labeling and prediction were favorably judged, as the opinion scores of stress and pauses were high above our level in the MOS test, among the top systems (see Figs. 7 and 8). Overall, it was positive to find out that the listeners were apparently able to analytically judge different aspects of speech; an important result considering prosody assessment of synthesis.

6. Conclusions

This year's challenge was very difficult, made even more challenging for us by the bad choice of training data. The noisy recordings were hard for our IAIF based vocoder, the large variability of styles for HMM training, and the paragraph length utterances for prosody prediction. Regardless, in light of the results, we were able to achieve a clean, intelligible voice with decent above average prosody characteristics. In the future, we will work on improving the robustness of the vocoder and the pulse library method, as well as prosody annotation with unsupervised methods. Also, retaining speaker characteristics, which could be our strong point with detailed source modeling, has not been very successful in these challenges, and should be improved. Finally, it could be that more interesting, focused research might have been made if the number of new topics were more limited. For example, studio quality audio book data could have provided enough challenge.

7. Acknowledgements

This research is supported by the EC FP7 project Simple4All (287678), Academy of Finland (1128204, 1218259, 121252, 135003 LASTU), MIDE UI-ART, and Tekes.

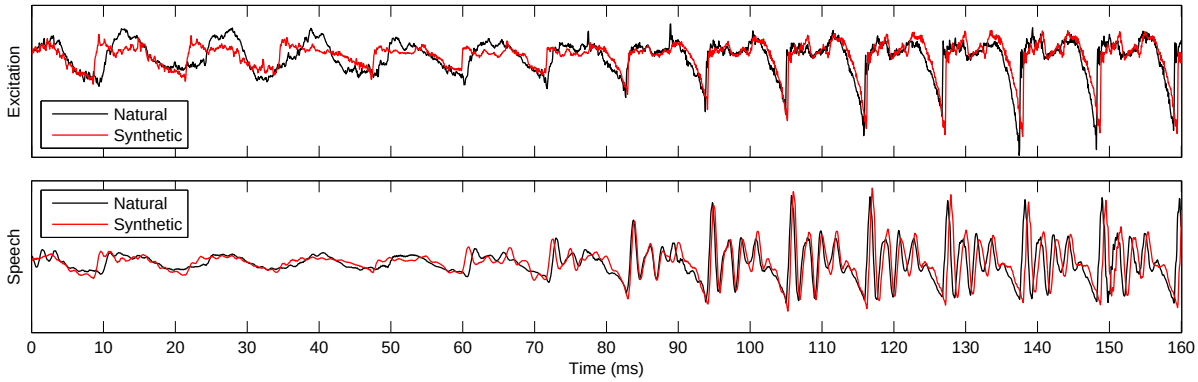

Figure 5: *Black line shows an estimated glottal flow signal (upper) of speech segment /ho/ (lower). Red line shows the corresponding synthetic glottal flow signal (upper) and speech segment (lower). The excitation is gradually changed from round pulses of breathy /h/ to sharp excitation peaks of modal /o/ due to the selection of appropriate pulses from the pulse library.*

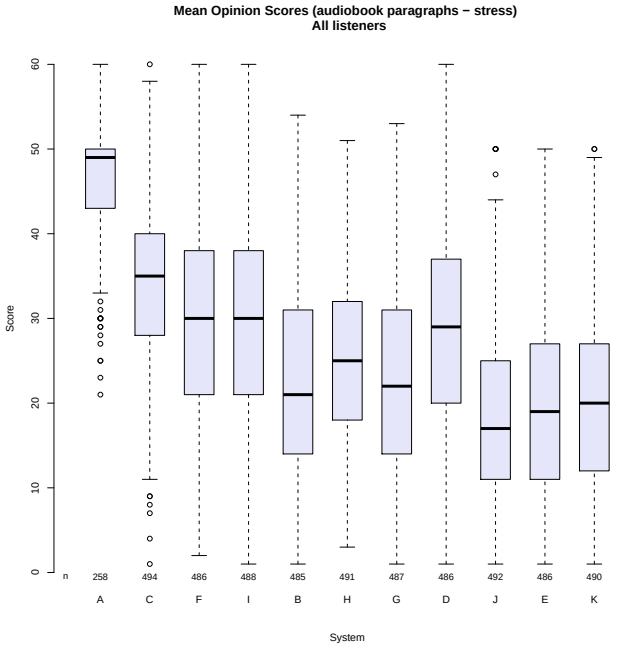

Figure 7: *Stress assignment results (D - GlottHMM).*

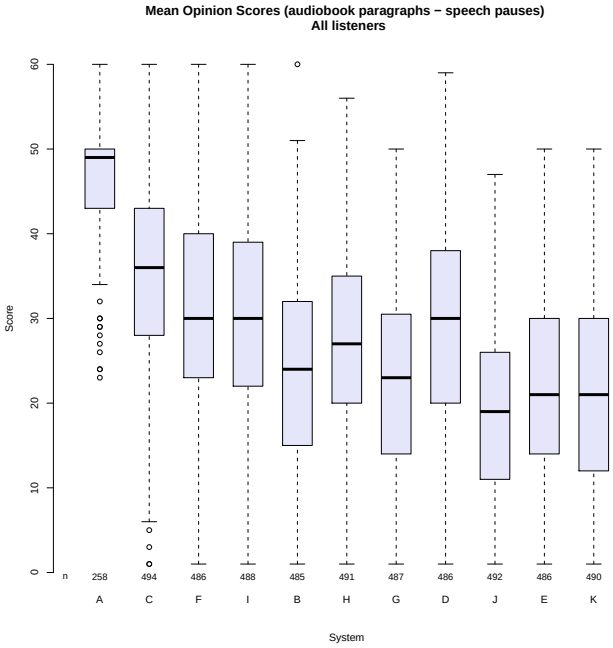

Figure 8: *Pause results (D - GlottHMM).*

8. References

[1] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.

[2] Suni, A., Raitio, T., Vainio, M. and Alku, P., "The GlottHMM speech synthesis entry for Blizzard Challenge 2010", The Blizzard Challenge workshop, 2010. Online: http://festvox.org/blizzard

[3] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis", ICASSP, 2011, pp. 4564–4567.

[4] TreeTagger. Online: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[5] Vainio, M., Suni, A. and Sirjola, P., "Accent and prominence in Finnish speech synthesis", Specom, 309–312, Oct. 2005.

[6] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Communication, 11(2–3):109–118, 1992.

[7] Alku, P, Tiitinen, H. and Näätänen, R., "A method for generating natural-sounding speech stimuli for cognitive brain research", Clinical Neurophysiology, 110:1329–1333, 1999.

[8] Qian, Y., Soong, F.K., Chen, Y., Chu, M., "An HMM-based Mandarin Chinese text-to-speech system", ISCSLP, 2006, pp. 223–232.

[9] Ling, Z.-H., Wu, Y., Wang, Y.-P., Qin, L. and Wang, R.-H., "USTC system for Blizzard Challenge 2006: An improved HMM-based speech synthesis method", The Blizzard Challenge Workshop, 2006. Online: http://festvox.org/blizzard

[10] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W. and Tokuda, K., "The HMM-based speech synthesis system (HTS) version 2.0", 6th ISCA SSW, pp. 294–299, 2007.

[11] Wu, Y.-J. and Wang, R.-H., "Minimum generation error training for HMM-based speech synthesis", ICASSP, pp. 89–92, 2006.

[12] Vainio, M. and Järvikivi, J., "Tonal features, intensity, and word order in the perception of prominence", J. of Phonetics, 34:319–342, 2006.