The GlottHMM Speech Synthesis Entry for Blizzard Challenge 2010

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

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

Abstract

This paper describes the GlottHMM speech synthesis entry for Blizzard Challenge 2010. GlottHMM is a hidden Markov model (HMM) based speech synthesis system that utilizes glottal inverse filtering for separating the vocal tract from the glottal source. The source and the filter characteristics are modeled separately in the framework of HMM. In the synthesis stage, natural glottal flow pulses are used to generate the excitation signal, and the excitation signal is further modified according to the desired voice source characteristics generated by the HMM. In order to prevent the over-smoothing of the vocal tract filter parameters, a new formant enhancement method is used to make the vocal tract resonances sharper. Finally, speech is synthesized by filtering the glottal excitation by the vocal tract filter.
Index Terms: speech synthesis, hidden Markov model, glottal inverse filtering

1. Introduction

GlottHMM text-to-speech (TTS) system [1, 2] is developed in a collaboration between Aalto University and University of Helsinki. In this entry, we have used our speech synthesis system that emphasizes the importance of the speech production mechanism, especially in terms of separating the two distinct parts of it: the glottal excitation and the vocal tract filter.

Compared to typical parametric synthesizers, our detailed model of the excitation should potentially allow for better control and production of prosody, speaker characteristics and speaking style, and this year's challenge had several interesting tasks to test this assumption. Thus, we participated in all tasks, except ES3 and MS1.

Comparison with other systems was an important motivation for participation, but the process of building the voices was also very useful in developing our system further. The EH1 and MH1 were by far the largest databases we have used, and required optimization of the training process. Additionally, these were our first serious attempts at building non-Finnish voices, and a prototype of English front-end was developed to test our prominence based prosody model [3] for a new language.

As this is our introduction to the challenge, we will first describe our synthesis system in some detail, followed by discussion on voice building and English front-end, and finally pick some selected results for analysis.

2. Overview of the system

Statistical parametric speech synthesis has recently become very popular due to its flexibility. However, the speech quality and naturalness of parametric speech synthesizers are usually inferior compared to state-of-the-art unit selection speech synthesis systems. This degradation is mainly caused by three factors: oversimplified vocoder techniques, acoustic modeling accuracy, and over-smoothing of the generated speech parameters [4]. The GlottHMM text-to-speech (TTS) system tries to overcome the problems especially with oversimplified vocoder techniques and the over-smoothing of the speech parameters.

The GlottHMM TTS system uses a vocoder technique that utilizes glottal inverse filtering [5]. In the parametrization stage, glottal inverse filtering is used to decompose the speech into the glottal source signal and the model of the vocal tract filter. This enables the separate analysis and modeling of the glottal source and the vocal tract filter, and thus the reconstruction of the excitation signal in the synthesis stage. The modeling of the voice source has been under intensive research recently, especially in HMM-based speech synthesis, and several techniques have been proposed to model the source signal [6, 7, 8, 9]. In synthesis stage, GlottHMM TTS system uses real glottal flow pulses extracted from natural speech for reconstructing the excitation signal, and the spectral characteristics of the excitation signal are further modified by filtering the signal with an adaptive IIR filter in order the preserve the desired voice quality.

The vocal tract filter is used to filter the excitation to generate speech. However, the over-smoothing of the speech parameters is especially severe perceptually when it affects the formant structure of speech. The over-smoothing excessively increases formant bandwidths, which results in unnatural perceptual quality of vowels. Thus, the GlottHMM system uses a new formant enhancement technique for sharpening the formants of the vocal tract spectrum [10]. The method is described in Section 2.3.

GlottHMM is built on a basic framework of an HMM-based speech synthesis system [11], but the parametrization and synthesis methods differ from conventional vocoders and are therefore explained in detail below

2.1. Speech parametrization

The flow chart of the speech parametrization algorithm is shown in Figure 1. First, the signal is high-pass filtered (cut-off frequency 70 Hz) in order to remove possible low-frequency fluctuation from the signal. This high-pass filtering is important since low-frequency fluctuation may create large errors to the glottal flow estimate, since glottal inverse filtering requires integration. The signal is then windowed with a rectangular window to 25-ms frames at 5-ms intervals and the speech features, presented in Table 1, are extracted from each frame. The order of all-pole modeling, depicted in Table 1, depends on the speaker, and usually lower order spectral models work well for female speakers while greater order spectral models yield better results for male speakers.

The log-energy of the windowed speech signal is evaluated, after which glottal inverse filtering is performed in order to estimate the glottal volume velocity waveform from the speech

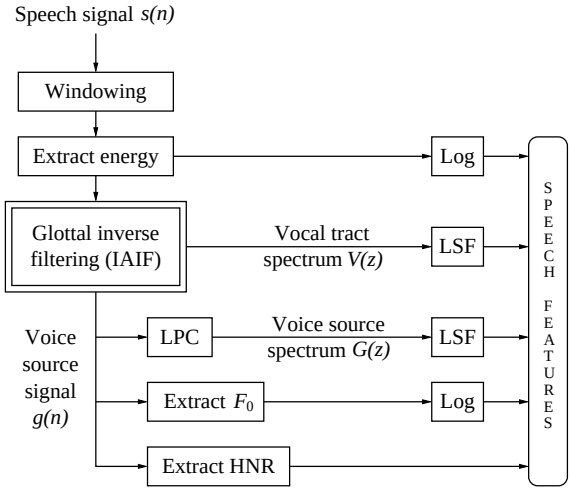

Figure 1: *Illustration of the parametrization stage. The speech signal $s(n)$ is decomposed into the glottal source signal $g(n)$ and the all-pole model of the vocal tract $V(z)$ using the IAIF method. The glottal source signal is further parametrized into the all-pole model of the voice source $G(z)$, the fundamental frequency F_0, and the harmonic-to-noise ratio (HNR). The obtained parameters are converted to a suitable representation for the HMM system.*

signal. An automatic glottal inverse filtering method, Iterative Adaptive Inverse Filtering (IAIF) [12, 13], is utilized. IAIF iteratively cancels the effects of the vocal tract and the lip radiation from the speech signal using all-pole modeling. Consequently, the outputs of the IAIF algorithm are the estimated glottal flow signal and the all-pole model of the vocal tract. In order to capture the variations in the glottal flow due to different phonation or speaking style, the spectral envelope of the glottal flow is further parametrized with linear predictive coding (LPC). This spectral model of the glottal excitation captures mainly the spectral tilt, but also the more detailed spectral structure of the source. The degree of the LPC analysis depends on the speakers, and analysis orders from five to ten have proven to work well.

The fundamental frequency is estimated from the glottal flow signal with the autocorrelation method. In order to evaluate the degree of voicing in the glottal flow signal, a harmonic-to-noise ratio (HNR) is determined 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 [14]. LPC models of the vocal tract and the voice source are further converted to line spectral frequencies (LSFs) [15], which provides stability [15] and low spectral distortion [16]. In case of unvoiced speech, conventional LPC is used to evaluate the spectral model of speech.

2.2. Synthesis

The flow chart of the synthesis stage is shown in Figure 2. The excitation signal consists of voiced and unvoiced sound sources. The basis of the voiced sound source is a glottal flow pulse extracted from a natural vowel. By interpolating the real glottal flow pulse according to F_0 and scaling in magnitude according to the energy measure, a pulse train comprising a series of individual glottal flow pulses is generated. In order to control the degree of voicing in the excitation, the amount of noise in the excitation is matched by manipulating the phase and magnitude of the spectrum of each pulse according to the harmonic-to-noise measure at each frequency band. Furthermore, the spectral tilt of each pulse is modified according to the all-pole spectrum generated by the HMM. This is achieved by filtering the pulse train with an adaptive IIR filter which flattens the spectrum of the pulse train and applies the desired spectrum. For voiced excitation, the lip radiation effect is modeled as a first-order differentiation operation. The unvoiced excitation is composed of white noise, whose gain is determined according to the energy measure generated by the HMM system. The vocal tract parameters are enhanced in order to alleviate for the over-smoothing, and the LSFs are then interpolated and converted to LPC coefficients, and used for filtering the excitation signal.

2.3. Alleviation for the over-smoothing

In order to alleviate for the over-smoothing of the vocal tract parameters, a new formant enhancement technique [10] is used to modify the LPC coefficients. The method is based on modifying the power spectrum of the all-pole model, and then re-evaluating LPC based on the modified power spectrum. The algorithm is described as follows. First, the power spectrum is evaluated from the LPC coefficients, for example, using the fast Fourier transform (FFT). This yields the spectral model of speech, which can be modified in order to enhance the over-smoothed formants. The modification procedure is based on additionally reducing the energy at the low-energy parts, i.e., the valleys of the spectrum. The reduction is performed by multiplying the low-energy regions with a small real-valued coefficient α, while the spectral peaks are left unmodified. The spectral peaks and the valleys are easily found from the smooth LPC envelope by searching for the zero-crossing points in the differentiated spectral envelope. After the additional reduction of the valleys, the modified power spectrum is inverse Fourier transformed into a new autocorrelation function, from which a new LPC model can be evaluated using the Yule-Walker equations. The new LPC model will most likely show sharper formants since LPC focuses on the spectral peaks in the frequency domain. For normal voices, $\alpha = 0.3$ was used for formant enhancement.

3. Building voices

In this chapter, we will describe the steps taken in building the English voices. Instead of using Festival tools, we made a decision to use this challenge as an opportunity to develop our own English front-end. We were especially interested in how our perceptual prominence based prosody modelling [3], developed for Finnish, would work for other languages. Unfortunately,

Table 1: Speech features and the number of parameters.

Feature	Parameters per frame
Fundamental frequency	1
Energy	1
Harmonic-to-noise ratio	5
Voice source spectr. (filter ord.)	10 (male), 7 (female)
Vocal tract spectr. (filter ord.)	30 (male), 20 (female)

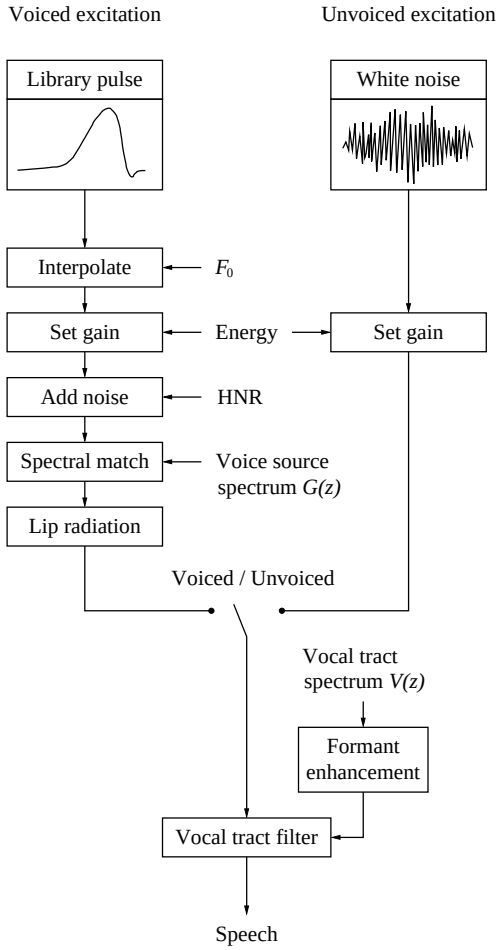

Figure 2: *Illustration of the synthesis stage. The basis of the voiced excitation signal is a library glottal flow pulse, which is modified according to the voice source parameters. Unvoiced excitation is composed of white noise. The excitation signals are combined and filtered with the vocal tract filter $V(z)$ to generate speech.*

due to time constraints, there are no proper evaluation results to report.

3.1. Data preparation

To cope with the difficulties of English orthography, we used the unilex lexicon (RP) as our pronunciation dictionary. To disambiguate between pronunciation variants, we enrich the texts with a part-of-speech tagger TagChunk[1]. Additional disambiguation on training data was performed with HTK by recognizing the most likely pronunciations with monophone models, trained from the provided monophone labels. The breaks were recognized similarly, allowing optional silences between words. At this point, we also discarded much of the training data, based on low probability scores that could indicate problematic utterances.

[1]http://www.cs.utah.edu/~hal/TagChunk/

3.2. Prosody modeling

In order to model sentence level prosody well, some form of phonological level has to be assumed. In English TTS systems, such as Festival, this level is usually represented by pitch accents. Each lexically stressed syllable can be either accented or not, with optional qualification of F_0 peak location and shape using ToBI transcription system. The accents typically fall on content words while function words are not often accented. The accent model has been successful in TTS because the presence or absence of accent can be annotated somewhat reliably based on text alone, and the prediction of accents in English TTS can be performed with good accuracy from shallow linguistic features, although the different ToBI accent classes are usually collapsed. This leaves a simple binary model, where the strength of accents can only be predicted indirectly, based on positional features and PoS information.

We have, however, used a model based on perceptual prominence, which could be regarded as a more fine grained accent decision, without the complexities of ToBI annotation. Additionally, while the pitch accent is, by definition, concerned with mainly one aspect of prosody, namely F_0, prominence can also be perceived based on duration, energy and spectral characteristics, and thus can affect the HMM clustering of all parameter streams favorably.

The prominences of the training corpora were annotated automatically, though in a supervised fashion. One of the authors annotated the prominences of 100 utterances from EH1 database, marking each syllable with a prominence value 0, 1, 2 or 3. These values correspond roughly to unaccented, nonnuclear accent, nuclear or other strong accent and emphatic or contrastive accent. The rest of the corpora were then tagged with a regression model. The features used to build the model were various syllable-level measurements (maximum, minimum, amplitude, etc.) of F_0, duration, energy and HNR. While this annotation method is prone to alignment and F_0 errors, the resulting tags are useful for HMM-synthesis training, as is evidenced by context-clustering trees where questions on prominence appear frequently and early on all streams. An example of prominence annotation and synthesis using correct prominence labels is shown in Figure 3, which gives indication on the accuracy of the synthesis of F_0, if the prominence labels were predicted correctly.

The prosodic boundaries were annotated in a similar fashion with strengths 1, 2, 3, and 4. The features used in annotation were the silence duration, and duration of syllables surrounding the silence. Also, a boundary of strength 1 was marked in a case of clearly lengthened syllable, even in absence of silence.

Annotated data was converted to our XML-based utterance structure, and contextual features were then extracted for full-context HMM-training. The features included typical positional and distance information of phones, syllables, words and phrases. Prominence features included syllable and word level information of prominence and prominence of surrounding units, distances to high prominence (2, 3) units and distances to focused (3) units. Notably, no PoS nor other high level linguistic information with indirect effect on phonetic realization was included in HMM-training.

3.3. Prosody prediction

To predict the prominences on TTS, we built a CART predictor on the previously automatically annotated EH1 corpus, using typical shallow features, such as PoS and unigram frequencies, with additional syntactic phrase information provided by

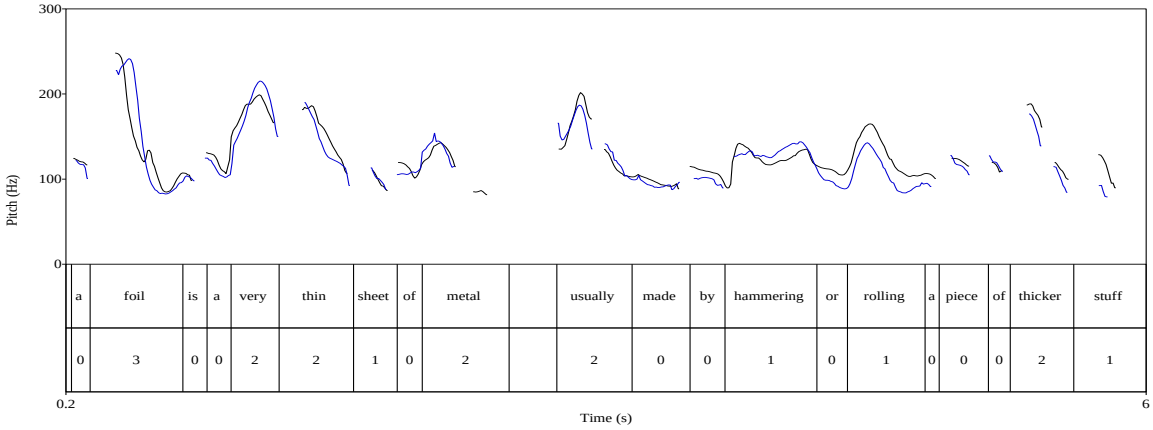

Figure 3: *Example of prominence annotation and synthesis of F_0 from EH1 database. The blue line shows the original F_0 contour and the grey line shows the contour produced by the synthesizer, given correct prominence labels.*

TagChunk. However, while the automatic annotation quality was good enough for HMM-training, the corpus was probably not large or consistent enough to produce a very good predicting model. The predictor failed to learn level 3 prominence completely, as the instances of emphasis in corpus were few and erratically annotated. Some additional rules based on lexical cues were written to alleviate the problem. For example, nouns following focus particles *even* and *only* were emphasized. Other rules were added to take into account some discourse-level phenomena, such as decreasing the prominence of given words in certain contexts.

Prosodic boundaries were predicted by rule, using punctuation, PoS and syntactic chunk information. Other front-end components, such as pre-processing and post-lexical rules, received very little attention.

3.4. Parametrization and HMM-training

HMM-training was prepared by parametrizing the speech files as described in Section 2.1. F_0 ranges of speakers were manually determined, and formant enhancement described in Section 2.3 was applied to vocal tract LSFs prior to training. English voices were trained with contextual labels of our own, and Mandarin voices with iFlytek labels.

HTS 2.1 was used to train context-dependent multi-stream and multi-space distribution (MSD) hidden semi-Markov models. The MSD structure was used for statistical modeling of the fundamental frequency. The rest of the features are modeled as continuous probability distribution (CD) streams. Other source features, the harmonic-to-noise ratio and the voice source spectrum would also be natural candidates for MSD modeling, but the results of such experiments have been mixed, so the submitted voices use CD streams. Only the vocal tract LSFs and F_0 features were considered during the alignment step of the parameter re-estimation; the weights of the other streams were set to zero. This is probably not optimal, but finding good stream weights on our multi-stream system is difficult.

Typical five-state left-to-right MSD-HSMMs were used for the submitted voices. Each state has a single Gaussian probability distribution function (pdf) with a diagonal covariance matrix as the state output pdf and a single Gaussian pdf with a scalar variance as the state duration pdf.

The training followed the de-facto standard procedure of HTS. Provided time alignments were used to get the initial estimates of monophone MSD-HSMMs. Monophones were then re-estimated five times, followed by conversion to full-context models, then re-estimation (1x), decision-tree based clustering, re-estimation (5x), re-estimation of untied models (1x), clustering and finally re-estimation (5x) of final models.

Each stream was clustered independently, as different contextual features affect different parameter types, and this way we can use the training data most efficiently. However, this method should probably be reconsidered, as our parameter types, particularly different source parameters are obviously not independent from each other, and the correlation between parameter types is largely lost in independent clustering. This is a possible cause for some artefacts present in submitted voices.

On the technical side, due to large number of parameters, our training process is somewhat more resource intensive than conventional HMM-synthesizers. Combined with large databases used in the challenge, this posed some difficulties on our training procedure, and the computer previously used for training. The problems were solved by moving the training platform to 64-bit Mac with 8 processor cores and 16GB of memory. With re-estimation and clustering steps performed in parallel, training of the largest voice, EH1, took six days real-time.

3.5. Resulting voices

For parameter generation from HMMs, we applied HTS_ENGINE, modified to use arbitrary number and type of streams, as defined in configuration file. Post-filtering and global variance expansion were applied to voices sparingly, as the formant enhancement prior to training provided clear formants. Resulting voices were considered to be of fine quality, except for a considerable number of voicing errors and fluctuating F_0, caused at least partly by careless tuning of F_0 estimation parameters, and possibly, problems with our labels. Lack of robustness on voicing boundaries have long been a weakness of our system, and the voicing errors in our system tend to cause particularly audible artefacts, due to inconsistencies between different source and vocal tract parameters.

Unfortunately, the front-end work left no time for solving these issues and re-training the large voices. We had to resort

to forced voicing decision in parameter generation; vowels and voiced consonants were set to be voiced and other phones unvoiced. This decision led to other problems, because, in reality, the voicing boundaries do not align perfectly on HMM-model boundaries.

3.6. Adapted Roger

Voice for task ES1 was trained by adaptation, using the HMMs of large EH1 data as the source voice and the 100 sentences from Roger as adaptation data. All parameter types were adapted, using CMLLR and SMAP algorithms provided by HTS. Because of difficulties in tuning the adaptation parameters, the resulting voice was quite unstable, and parameters had to be smoothed in synthesis.

3.7. Speech in noise tasks

This year's Blizzard Challenge included tasks to build voices suitable to be heard in the presence of additive noise. For these tasks, the voices trained for tasks EH1 and MH1 were used as bases. Taking advantage of the flexibility of our system, we tried to model several aspects of what is known of actual people speaking in the presence of noise, or the Lombard effect [17]. Since naturalness and speaker similarity were not considered in the evaluation, large modifications were possible.

Firstly, to model careful articulation, we increased the prominence of all stressed syllables in content words and applied more post-filtering to produce clearer formant structure. In addition, the post-filtering was used only during the open phase of the glottal cycle. This procedure tries to imitate the physiology of the speech production mechanism, where the formants are actually stronger while the glottis is open, providing a longer acoustic tube. Also, the rate of speech was lowered by a factor of 0.9 in parameter generation.

Secondly, to facilitate following of the speech flow, pitch was raised by a factor of 1.4 with compressed range. Vocal tract length was also shortened slightly to match the raised pitch and raise formant frequencies, as observed in real Lombard speech. Further, the intra-utterance silences were replaced with boundaries with pre-boundary lengthening only to provide continuous speech flow.

Thirdly, to model the speakers' efforts to make the speech more audible, the harmonic structure of the glottal pulse was modified to be sharper, modeling the decreased spectral tilt in the voice source while loud voices are produced. The spectral tilt was halved from the normal spectral tilt. This has an effect of concentrating most of the energy on formant frequencies, the most sensitive frequencies of human auditory system. The resulting signal waveform was then compressed in order to make the loudness of the speech as high and uniform as possible.

Subjectively, the resulting voice sounded very much like air force radio communication. Clear improvement on intelligibility compared to baseline voices was confirmed by authors in presence of babble noise and white noise.

4. Results and discussion

4.1. English

This year's challenge had many high quality entries on English tasks, as evidenced by the general rank drop of the HTS baseline systems, compared to last year. Given the problems in voice preparation, the results on English hub tasks were as expected; our MOS scores were consistently higher than STRAIGHT-based HTS baseline systems, though not significantly so in all tasks, but we could not compete with the best voices.

Among quite similar systems, the frequent artefacts and unstable F_0 contour probably affected our scores severely. The intelligibility scores, typically a strong point for parametric synthesis, were also only average. This can be attributed to voicing problems, which affected the stop consonants, and possible omissions and bugs in the pronunciation component of our front-end. Overall, the effect of the described front-end and prosody modeling can not be confirmed either way. As the test sentences were short and quite uniform in structure, the detailed prosody modeling was probably not particularly advantageous in this year's challenge.

4.2. Mandarin

We were pleased to note that our system performed well on the Mandarin tasks, especially as high pitched female voices are difficult for the IAIF method. Our system ranked among the best on MOS on task MH1, as shown in Figure 4, with our system marked with 'N'. On the other hand, the similarity scores were quite bad for the Mandarin voices, even compared to HTS2005 baseline. The reasons for this are not completely clear to us. Maybe our experimental formant enhancement methods or the glottal pulses extracted from different speakers change the voice characteristics more than we had expected. On intelligibility tests, only the original speaker ranked significantly higher than our system, which further suggests that the English results were affected by temporary problems.

4.3. Speech in noise

The tasks ES2 and MS2 were clearly suitable for our synthesis approach. The high modifiability of our system ensured that our voices had the lowest word error rates by a clear margin, even compared to the original speakers (Pairwise Wilcoxon signed rank tests shows significance at $p < 0.01$). It must be said though, that this comparison was unfair to original speakers, as their speech was recorded in silent conditions. The results of the English voice are shown in Figure 5.

5. Conclusions and future work

In this paper, we have described the GlottHMM speech synthesis system used for Blizzard challenge 2010. Our previous work has been done almost exclusively on Finnish synthesis, so it was very valuable to get the feedback of our system's status in the whole speech synthesis community. Participation provided us with important information on the strengths and weaknesses of our system. The results of the hub tasks showed that we have to work on the robustness and speaker similarity issues of our system. On the other hand, we got support on the relevance of our physiologically inspired synthesis approach on the good results of the speech in noise tasks.

6. Acknowledgements

The research in this paper supported by Nokia, the Academy of Finland (projects 135003, 107606, 1128204, 1218259, research programme LASTU), and MIDE UI-ART.

7. References

[1] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "HMM-based Finnish text-to-speech system utilizing glottal inverse filtering", Proc. Interspeech, 2008.

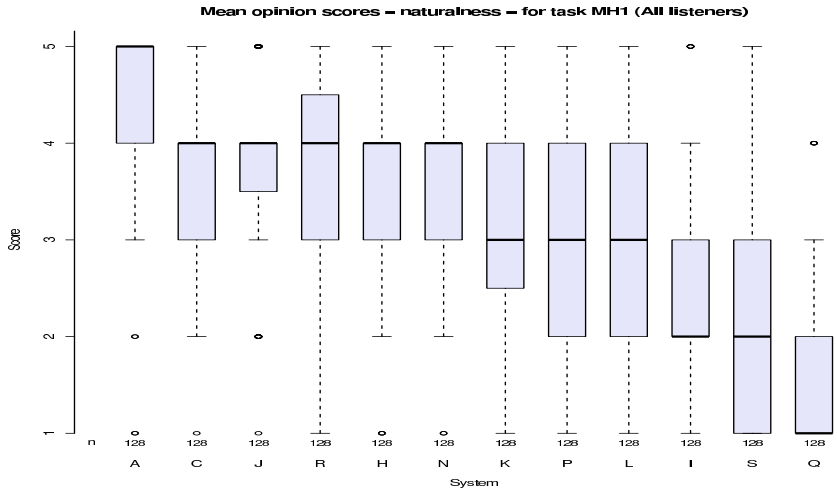

Figure 4: *Mean opinion scores on Mandarin hub task 1.*

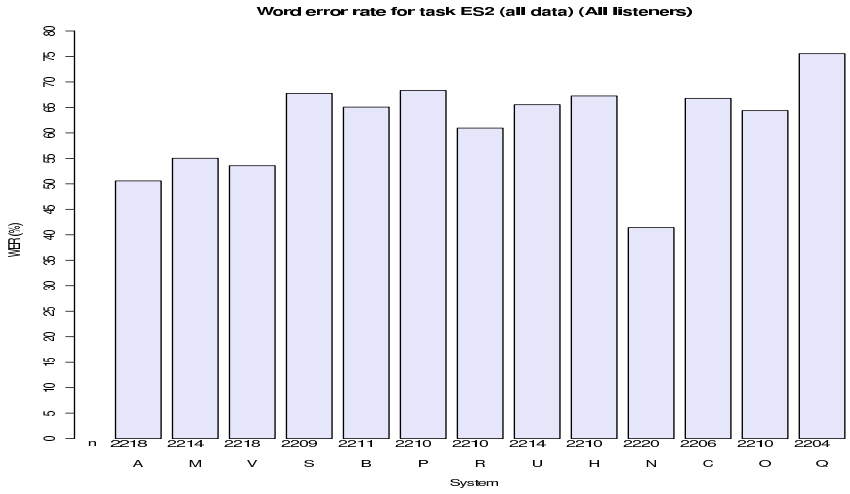

Figure 5: *Word error rates (WER) for the English speech in noise task.*

[2] 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. Audio, Speech, and Language Processing, (in press).

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

[4] Zen, H., Tokuda, K. and Black, A. W., "Statistical parametric speech synthesis", Speech Commun., 51(11):1039–1064, 2009.

[5] Miller, R. L., "Nature of the vocal cord wave", J. Acoust. Soc. Am., 31(6):667–677, Jun. 1959.

[6] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T., "Mixed excitation for HMM-based speech synthesis", Proc. Eurospeech, pp. 2259–2262, 2001.

[7] Maia, R., Toda, T., Zen, H., Nankaku, Y. and Tokuda, K., "An excitation model for HMM-based speech synthesis based on residual modeling", Sixth ISCA Workshop on Speech Synthesis, Aug. 2007.

[8] Kim, S. J., and Hahn, M., "Two-band excitation for HMM-based speech synthesis", IEICE Trans. Inf. & Syst., vol. E90-D, 2007.

[9] Fant, G., Liljencrants, J. and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, 4:1–13, 1985.

[10] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P., "Comparison of formant enhancement methods for HMM-based speech synthesis", accepted for publication in Seventh ISCA Workshop on Speech Synthesis, Kyoto, Japan, 2010.

[11] 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", Sixth ISCA Workshop on Speech Synthesis, 294–299, Aug. 2007.

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

[13] 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.

[14] Moore, B. C. J. and Glasberg, B. R., "A revision of Zwicker's loudness model", ACTA Acustica, 82:335–345, 1996.

[15] Soong, F. K. and Juang, B.-H., "Line spectrum pair (LSP) and speech data compression", Proc. ICASSP, 9:37–40, 1984.

[16] Paliwal, K. and Kleijn, W., "Quantization of LPC parameters", Speech Coding and Synthesis, W. Kleijn and K. Paliwal, Eds. Elsevier, ch. 12, 1995.

[17] Van Summers, W., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I. and Stoke, M. A., "Effects of noise on speech production: Acoustic and perceptual analyses" J. Acoust. Soc. Am., 84(3):917–928, 1988.