UCD Blizzard Challenge 2011 Entry

Peter Cahill, Udochukwu Ogbureke, João Cabral, Eva Szekely,
Mohamed Abou-Zleikha, Zeeshan Ahmed, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics, University College Dublin, Ireland.
peter.cahill@ucd.ie, kalu@ucdconnect.ie, joao.cabral@ucd.ie, eva.szekely@ucdconnect.ie,
mohamed.abou-zleikha@ucdconnect.ie, zeeshan.ahmed@ucdconnect.ie, julie.berndsen@ucd.ie

Abstract

This paper gives an overview of the UCD Blizzard Challenge 2011 entry. The entry is a unit selection synthesiser that uses hidden Markov models for prosodic modelling. The evaluation consisted of synthesising 2213 sentences from a high quality 15 hour dataset provided by Lessac Technologies. Results are analysed within the context of other systems and the future work for the system is discussed.

Index Terms: blizzard challenge, speech synthesis, unit selection, duration modelling

1. Introduction

This paper presents an overview of the UCD Blizzard Challenge 2011 system and the evaluation results. The motivation for the entry was to obtain an independent comparison with other systems.

The UCD speech synthesis system was designed to be a research tool. A lot of design effort was focused on a highly modular architecture that was trivial to utilise features in other tools. While the core of the system was developed in C#, the synthesiser invokes functionality from external tools such as MATLAB, shell scripts, HTK, etc. using the Muse platform [1].

Alternative algorithms are straightforward to add, particularly for common tasks, such as forced-alignment, distance measurements, grapheme-to-phoneme, prosodic features, etc. It is possible to build HTS voices directly from unit selection voices (using the same annotations), where we have begun to experiment with hybrid systems.

2. System

2.1. Overview

The configuration of the system for the Blizzard Challenge 2011 is a generic unit selection system. The system uses diphone units, where continuous units have a join cost of 0. The join feature vectors consisted of 13 MFCC parameters, along with first and second order regression parameters, and log f0. The target feature vectors comprised of phonetic contexts, duration, and f0. Feature vector distances were measured with a weighted Euclidean distance function, where all weight values were set manually.

Target utterances were estimated purely from models that were trained from the voice data (i.e. no external data was used), with the exception of the Unilex dictionary [2] which was used for grapheme-to-phoneme conversion.

The decoder used was a Viterbi decoder (i.e. no pruning in search path). The join function was a basic raw join function that would join segments at 10ms boundaries with a basic optimal coupling technique which encourages joins at zero crossings. The join function does not attempt to modify pitchmarks.

Statistical models were used to model duration and F0. One of the more unique aspects of the system is the approach to duration modelling. This is discussed in the following section.

2.2. Duration Modelling

The duration modelling is one of the unique features of the system. The duration is modelled using hidden Markov models (HMMs). It is an alternative to the hidden semi-Markov models (HSMMs) commonly used for explicit duration modelling in HMM-based speech synthesis [3]. The modelling is a two-step process as shown in Figure 1. The first step in this process is state level monophone alignment and quantisation (conversion into number of frames). In the second step, HMMs are trained whereby the observations are the number of frames in each state and the hidden states are the phones. This enables the duration of each state (the number of frames) to be generated from the trained HMMs. The duration of each phone is the sum of the state durations.

2.2.1. Duration training

The topology of the alignment model is shown in the upper part of Figure 2. This is a 5-state HMM, as normally used for acoustic modelling in speech recognition and synthesis. The topology for the explicit duration modelling is shown in the lower part of Figure 2 as a 5-state HMM in which the state emission probability of state j, namely $b_j(o)$, is modelled by a Gaussian distribution with mean ξ_j and variance σ_j. The first training step is the estimation of the observation. This involves training of monophone models, followed by state level monophone alignment. The duration of each state is defined by the number of speech frames and is obtained by dividing the duration, in milliseconds, by the frame rate. The second training step involves the training of explicit duration models. A different HMM with the same number of states as the alignment model is trained whereby the observations are the number of frames obtained from the first step and phones are the hidden states. The training involves training of monophones, followed by full context-dependent models. This is followed by context clustering of duration based on phonetic and prosodic contexts and retraining of clustered models. This duration modelling approach is simpler as it uses continuous HMMs and produces comparable results for speech synthesis. More details about the model and the evaluation will be published in the near future as an evaluation is currently underway.

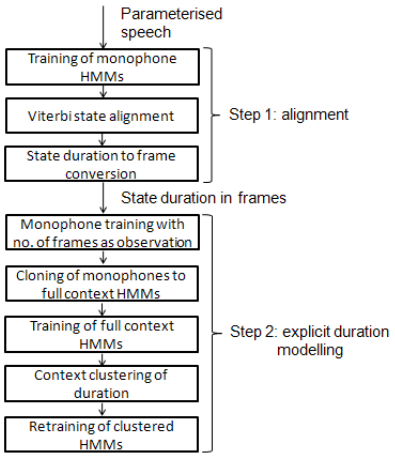

Figure 1: Training stages for the duration model.

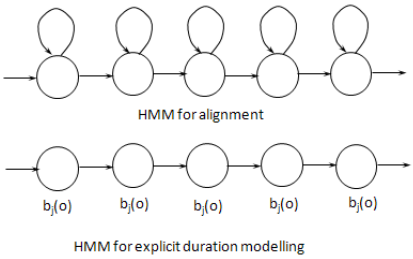

Figure 2: The prototype HMM of the duration model showing state observation distributions.

2.2.2. *Duration generation*

Given an input phone sequence with phonetic and prosodic context information, durations are generated from HMMs by maximising (1) in order to determine state durations for a sentence HMM [4].

$$log\ P\left(O|q,\lambda\right) = -\frac{1}{2}O^{'}\sigma^{-1}O + O^{'}\sigma^{-1}\xi + K, \quad (1)$$

where q is the given state sequence and K is a constant. Equation 1 is maximised when $O = \xi$, that is, the set of duration parameters that maximises the equation becomes the sequence of the mean vector [4]. The total duration T for a sentence HMM is the summation of the means of the states,

$$T = \beta \sum_{j=1}^{J} \xi_j, \quad (2)$$

where J is the total number of states in the concatenated HMMs and β is a positive scaling factor that controls the speaking rate. Since T is a function of β, the speaking rate is controlled by increasing or decreasing the number of frames in the utterance.

2.3. F0 modelling

Fundamental frequency (F0) models were trained using HTS [3]. The F0 contour is a mixture of values in the voiced and unvoiced region of the speech signal. F0 is modelled with HMMs based on multi-space distribution (MSD) [5]. F0 modelling based on MSD is the standard method used to model F0. MSD

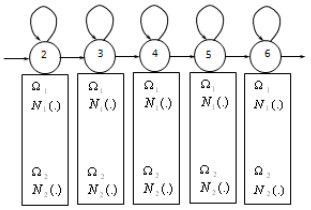

Figure 3: A 5-state MSD-HMM in which each space is associated with a probability distributions.

can simultaneously model a mixture of different distributions. Each distribution type is modelled in each of the spaces, thus the name MSD. A mixed distribution with G spaces can be represented by

$$\Omega = \bigcup_{g=1}^{G} \Omega_g \quad (3)$$

$$\sum_{g=1}^{G} W_g = 1, \quad (4)$$

where Ω represents the set of spaces and G the number of spaces. If Ω is a continuous distribution, then G represents the number of mixture components in the distribution and for a space made of discrete distribution, G represents the codebook size.

The F0 contour is a sequence of continuous values in the voiced region and discrete values in the unvoiced region. Thus G is equal to 2,

$$\Omega = \{\Omega_1, \Omega_2\}, \quad (5)$$

where Ω_1, Ω_2 represent the voiced and unvoiced spaces respectively. Figure 3 shows a 5-state MSD-HMM used for modelling F0 in which each state is associated with two probability distribution functions.

2.3.1. *F0 generation*

F0 is generated from a MSD-HMM as explained in [4] by the following equation:

$$log\ P\left(O|q,\lambda\right) = -\frac{1}{2}O^{'}U^{-1}O + O^{'}U^{-1}M + K, \quad (6)$$

where O is the F0 (observations); q is the state sequence given by the duration model; U and M are the covariance matrix and mean respectively, while K is a constant. The weight is used for making the voiced/unvoiced decision in each state. Typically, if the weight of the unvoiced distribution is less than 0.5, the state is assumed to be voiced. The F0 for each phone is the mean of the state F0.

3. Participation

3.1. Voice Building

The automatic voice building tool was developed with the ambition of consistently producing high quality voices automatically. In practice, current experiments suggest that the tool has become optimised for some languages (including English) but the longer term focus of the tool is to be consistent across languages. At a minimum, voices can be built from:

- A dictionary
- A collection of audio recordings
- Orthographic transcriptions that correspond to the audio recordings

The voice building process does not require any other information (manual alignments, annotations, etc.).

The voice building process involves calculating the target utterances for each utterance in a spoken corpus. This process includes estimating phonemes, syllables, phrases, and part-of-speech tags. These target utterances are then force-aligned to obtain appropriate temporal end points on the phone level. Acoustic and target features are then extracted from the data, and are used for the join and target costs respectively.

The Blizzard Challenge 2011 entry used treetagger [6] to calculate part-of-speech tags and phrase chunks. Force-alignment was performed using the embedded Baum-Welch hidden Markov model algorithm in HTK [7]. F0 was estimated using ESPS get_f0. Mel-cepstral frequency coefficients were calculated using the HCopy tool in HTK. The general American dialect of the Unilex dictionary was used for grapheme-to-phoneme conversion.

3.1.1. Lessemes

One of the interesting aspects of the Blizzard Challenge 2011 was the supplied annotations, known as *lessemes*. Traditional speech synthesis systems use phone sets with typically 40-50 annotation labels per language. The motivation for this is to classify the sounds in the dataset into groups which do not contain too much variation, while at the same time the groups are not too sparse. Lesseme labels are far more detailed that traditional phone sets, where there are several hundred labels for English. These labels include a phone classification, similar to a traditional phone set, and some of the following details (depending on if the phone is a vowel or consonant):

- dot-level
- stress
- inflection
- playability

The meaning of each of these is described in the provided Lesseme documentation.

Within the context of a unit selection synthesiser, this additional information can be incorporated in at least 2 ways:

1. Use the lesseme labels in place of the phone labels. This is the easiest way to incorporate lessemes. However, informal listening tests highlighted that the quality is low when this approach is used. This is because even though there might be appropriate units in the voice, if their lesseme label is not identical they will not be considered. In a diphone unit selection system this becomes a more significant problem, as if each unit label is of the form *unita-unitb*, then in the case of lessemes, each diphone will represent the transition from one lesseme to another. This results in many units having too few occurrences to perform high quality synthesis.
2. Use the phone label part of the lesseme to create units. This is similar to a traditional unit selection system where 40-50 phone labels are used to classify the units. The additional details that are represented in the lesseme label can be added as features in either the target or join cost functions. In practice, this method avoids the shortage of units problem however other problems are introduced:
   (a) Consonant and vowels can have the same phone label (e.g. there is a [Y] vowel and a [Y] consonant).
   (b) Weights need to be assigned to each feature added.

Informal listening tests highlighted that neither of these methods performed as well as the default, fully automatic voice building tool. The final submission therefore was not using the lesseme information (either labels or temporal endpoints).

3.1.2. Analysis

The entire voice data consisted of 15 hours and 6 minutes of audio (including silence at the start and end of the recordings). This data is split into 12095 utterances, of which 909 utterances were held-out of the voice. The motivation for pruning some utterances was that they were not in the Unilex dictionary and therefore may have incorrect phonemes or syllables estimated. In many cases, such words will be proper nouns, which often are incorrectly converted using grapheme-to-phoneme methods. The built voice contained 677,830 diphone units.

The final submission included sentences from the Blizzard Challenges 2009, 2010, and 2011. In total 2213 utterances needed to be synthesised. The 2213 synthesised utterances had a total duration of 2 hours and 37 minutes where only 1 hour and 28 minutes of voice data was selected for the target utterances. The most popular diphone, k-s, was selected 84 times. This is likely to be because there are not that may k-s diphones in the voice data, as diphones with vowels are likely to be more frequent. The synthesis task involved using 9251 of the utterances in the voice.

4. Evaluation

4.1. Overview

The format of the evaluation conducted on-line was similar to that used in recent years. It was divided in several parts for evaluating speech quality in terms of:

- Similarity with original speaker
- Mean opinion scores (speech naturalness)
- Intelligibility

Both the similarity and mean opinion scores (MOS) parts included two tasks for evaluating news and novel sentences, respectively. There was also a MOS task that included reportorial sentences (this sentence type has not been used in the Blizzard Challenge of previous years). Another difference to previous years was an additional intelligibility task using addresses, besides the typical intelligibility tasks using semantically unpredictable sentences (SUS).

The speech database released for the Blizzard Challenge contained both speech sampled at 16 and 48 kHz. It was possible to submit synthetic speech at 16 and 48 kHz (the released speech database included speech sampled at both rates). We have chosen to submit synthetic speech sampled at 16 kHz only because this is the sampling rate we usually work with, even though the voice building process in our system is the same for the two sampling rates.

The evaluation included 13 systems. The UCD system is represented by the letter “J”. The system A is the reference

natural speech, system B is a benchmark unit-selection voice built using Festival, system C is a speaker-dependent HMM-based voice and system D is the same as system C, except using 48 kHz sample rate data.

4.2. Similarity

The similarity results for all systems are given in Figures 4 and 5, for the novel and news domains respectively. Table 1 shows the results of the Wilcoxon's signed rank tests, which indicate if the difference between two systems is statistically significant. In Figures 4 and 5, the systems are ordered in descending order of the MOS means, although the *ordering is not a ranking* (the means are used to make the graphs more intuitive and should not be used to draw any conclusion from). The value of n in Figure 4 indicates the number of data points, which is the same for all systems. The median is represented by a solid bar across a box showing the quartiles. Whiskers extend to 1.5 times the interquartile range and outliers beyond this are represented as circles.

In general, the similarity to the original speaker is still far from that of natural speech. Our system (letter "J") obtained better similarity results for news than novel sentences. For the news domain, it was significantly better than systems F, D, and I, while it was only significantly worse than systems G and E. The results for the novel domain are somehow disappointing because the group of systems which were significantly better than system J is larger and includes the benchmark unit-selection system (letter "B"). This result could be explained by the fact that one of the weakest points of our system is the prosody modelling (it's under development stage) and it is expected that the novel sentences are richer in prosody than the news.

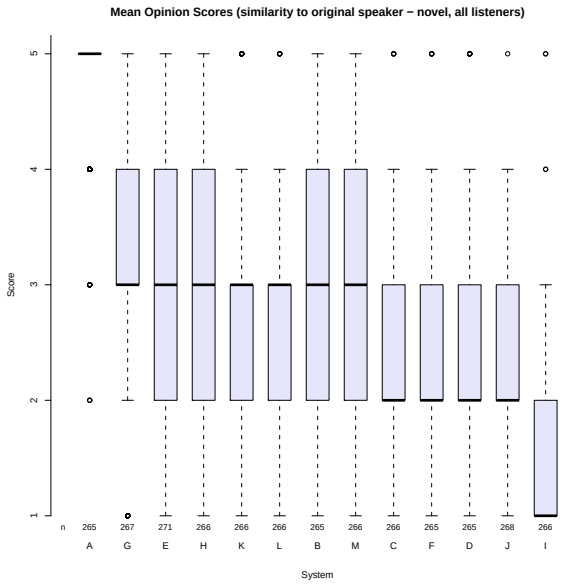

Figure 4: Scores of similarity for novel sentences.

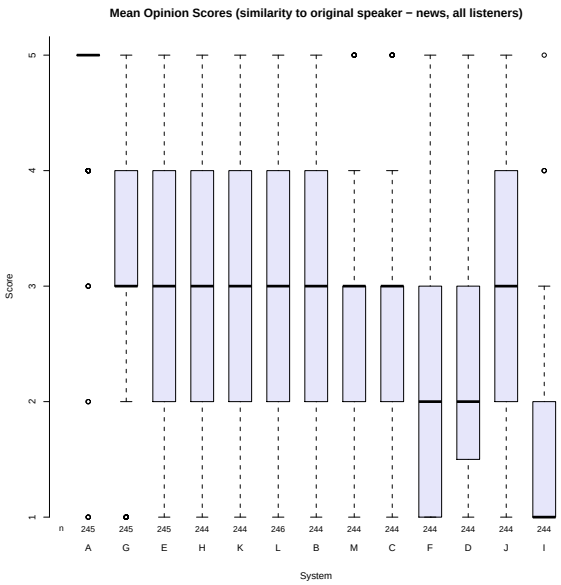

Figure 5: Scores of similarity for news sentences.

Data	Systems
All data	C, F, D
Novel	C, F, D
News	H, K, L, B, M, C

Table 1: Results of pairwise Wilcoxon signed rank tests for similarity scores. The system identifiers listed here are those that are **not** significantly different from system J.

4.3. Naturalness

Figure 6 shows the boxplot of the MOS and Table 2 shows the significant differences between the systems, for all data and listeners. The MOS are represented using a boxplot which is the same type as that used to show the similarity scores. From these results, systems A (natural speech), G, E, H, K, and L are significantly better in terms of naturalness than system J. The remaining systems are not statistically different from system J (including the benchmarks systems), with the exception of system "I". In general, the statistical differences between system J and the other systems for the novel and news domains are similar to those obtained for all data. The results for the reportorial sentences also show the same trend but system J is significantly worse than the benchmark HMM-based voices in this case.

Data	Systems
All data	B, M, C, F, D
Novel	B, M, C, F, D
News	B, C, F, D

Table 2: Results of pairwise Wilcoxon signed rank tests for MOS. The system identifiers listed here are those that are **not** significantly different from system J.

4.4. Intelligibility

Figure 7 shows the mean values of the word error rate (WER), for all systems and address data. The results of the Wilcoxon's signed rank tests indicated that there is not any significant dif-

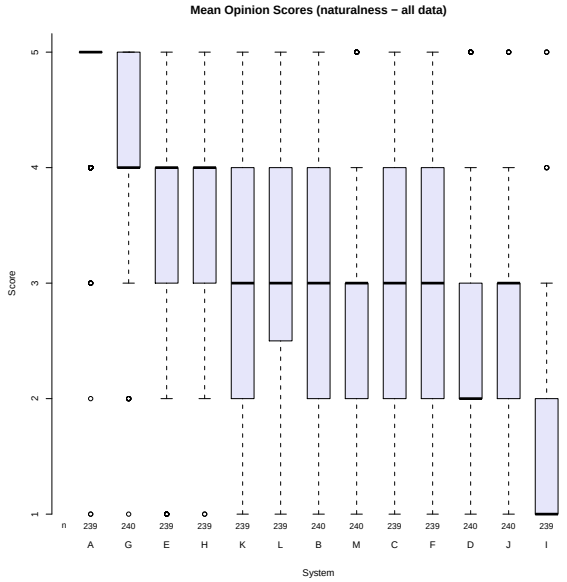

Figure 6: Naturalness scores (MOS) for all data.

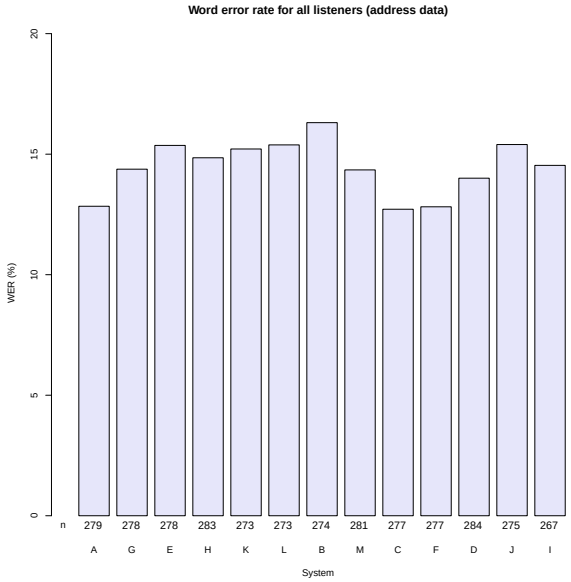

Figure 7: Mean WER for address sentences.

ference between systems. These results mean that all speech synthesis systems performed very well in this task, since they were not statistically different from natural speech. Figure 8 shows the WER results for the SUS data. In this case, all the speech synthesis systems are also not significantly different between each other, but natural speech is significantly different from every other system and obtained lower WER. Note that results computed for UK paid students and native English listeners were similar to the previous results in terms of the Wilcoxon's signed rank tests. However, for these two listener types the means of the WER for all systems were significant lower than considering all listeners.

4.5. Discussion

The evaluation results highlighted the current strengths and weaknesses the system. Prosody modelling is still a weakpoint in comparison to other systems. An in-depth study into both statistical and examplar methods are likely to help identify which area should be focused on in the near future. However, prosodic modelling is not limited to the underlying modelling algorithm. Quality prosodic modelling for the Blizzard Challenge should be able to apply speaking styles where appropriate. For example, alternative prosody models could be used for the different domains.

The next most significant weakness in the system is the current join function. The technique used in the system is very basic, and as such can often distort the synthesised signal. High-frequency clicks are frequently audible in the results because of the join function. Preliminary experiments were done using a more sophisticated method which did not create as much distortion, however, due to time constraints it was not possible to include it in the evaluation system.

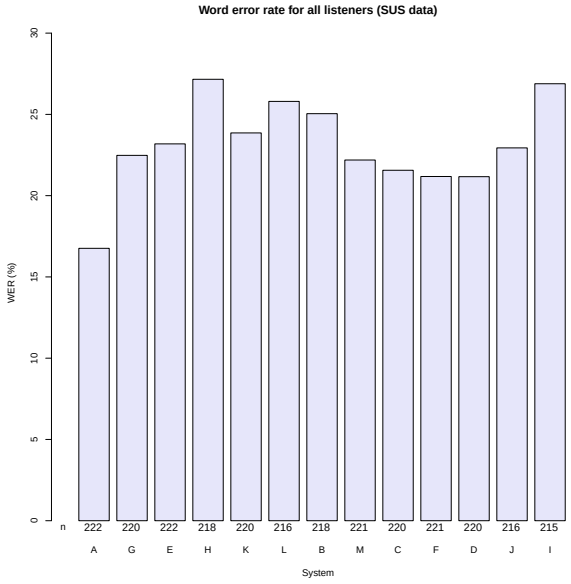

Figure 8: Mean WER for SUS.

5. Conclusions

This paper presents the UCD Blizzard Challenge 2011 entry. The system entered was a unit selection synthesiser that used HMMs for duration and MSD-HMMs (via HTS) for F0 modelling. The 15 hours and 6 minutes data set resulted in a voice with 677,830 diphones from our fully automatic voice building tool.

The results highlighted that in terms of intelligibility, the system is not significantly different from natural speech when reading addresses. However, it is significantly different when

reading sentences in the SUS domain. In terms of similarity and naturalness, the system performed best at reading news. The difference in results for news and other categories suggests that our current prosodic modelling is perhaps tuned for reading the news. In terms of similarity and naturalness, the UCD system did not perform as well as some systems but it was generally comparable to the benchmark systems and a number of other systems.

In future work, we intend to improve the prosodic modelling components and also integrate a new join function.

Voice style features are a technique that we have made some recent progress at which are likely to improve naturalness. A paper at Interspeech 2011 discusses the current approach [8]. This technique was specifically developed for audiobooks so it will be relevant to evaluate it in the Blizzard Challenge 2012.

Within the context of the Blizzard Challenge 2012, we will incorporate additional modifications which are appropriate for the audiobook-based datasets. Participation in the Blizzard Challenge helped identify the strengths and weaknesses of the system.

6. Acknowledgements

This work is supported by Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie). The authors would also like to express their thanks to Lessac Technologies for providing the voice data, and to the Blizzard Challenge organisers for their efforts.

7. References

[1] P. Cahill and J. Carson-Berndsen, "Muse: An open source speech technology research platform."

[2] S. Fitt, "Documentation and user guide to unisyn lexicon and post-lexical rules," *Center for Speech Technology Research, University of Edinburgh, Tech. Rep*, 2000.

[3] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-markov model based speech synthesis," in *Proceedings of ICSLP*, vol. 2, 2004, pp. 1397–1400.

[4] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, vol. 3, 2000.

[5] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution hmm," *IEICE Transactions on Information and Systems*, vol. 85, no. 3, pp. 455–464, 2002.

[6] H. Schmid, "Treetagger-a language independent part-of-speech tagger," *Institut f*
"ur Maschinelle Sprachverarbeitung, Universit
"at Stuttgart, p. 43, 1995.

[7] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*. Entropic Research Laboratory, Cambridge, England, 1997.

[8] E. Szekely, J. Cabral, P. Cahill, D. Aioanei, and J. Carson-Berndsen, "Clustering reading styles in audio-book data using voice quality parameters," in *Interspeech*, 2011.