The Toshiba entry for the 2007 Blizzard Challenge

Sabine Buchholz [1], Norbert Braunschweiler [1], Masahiro Morita [2], Gabe Webster [1]

[1] Speech Technology Group, Cambridge Research Lab, Toshiba Research Europe Ltd., Cambridge, UK

[2] Multimedia Lab, Toshiba Corporate Research and Development Center, Kawasaki, Japan

{sabine.buchholz,norbert.braunschweiler,gabe.webster}@crl.toshiba.co.uk

masahiro.morita@toshiba.co.jp

Abstract

This paper describes the system with which we took part in the Blizzard Challenge for the first time. It describes how we created our own annotation from scratch and introduces the various system components, in particular the back-end. Results show that our system achieves excellent intelligibility, in particular with less data (e.g. the lowest word error rate for voice C [1]), and reasonable naturalness (MOS of 3.0, 2.9 and 2.9 for voice A, B and C, respectively). They also reveal that our simple approach of randomly selecting sentences for voice C worked well.

1. Introduction

Members of the Multi-Media Lab and the Speech Technology Group of Toshiba jointly work on research and development of TTS technology for Asian and European languages for embedded applications. For these purposes, a system with a small memory footprint working in real-time is imperative, and intelligibility is often crucial for safety.

Typically, our TTS voices are built from high-quality in-house voice databases with carefully manually corrected annotations. Recently however, we started work into rapid development of voices, for which manual annotation has to be replaced by automatic annotation methods. The 2007 Blizzard Challenge (see [1] for details) was a good opportunity to test this line of work on a completely new voice database, under tight time constraints, and we therefore decided to take part.

The tasks for this 2007 Challenge, the third of its kind, were to build three synthetic voices from an 8-hour recording of one American English male speaker, containing sentences extracted from novels (the Arctic subset [1]), news, and travel conversation (the BTEC subset) [3]:

- voice A, using the full corpus (6,579 utterances)
- voice B, using only the Arctic subset (1032 utterances)
- voice C, using a subset of our choice with a total duration not exceeeding that of the Arctic subset (2,914 seconds, i.e nearly 50 minutes).

In the remainder of this paper, we describe the Toshiba TTS Research system used for this Challenge (Section 2), the preparation of the voice corpus (Section 3), system training and parameter optimization (Section 4) and work done for the three separate voices (Section 5). Finally we discuss results (Section 6) and conclusions and future work (Section 7).

[1] although not statistically significant from some other systems

2. The Toshiba TTS Research system

The Toshiba TTS Research system is a half phone-based unit-selection system using explicit prosody prediction and modification. More details about the system used for European languages are given in the following two subsections.

2.1. Front-end

First, the input text is split into sentences. For the Blizzard data, this was only needed for some of the BTEC utterances that consisted of more than one sentence. Next, sentences are split into tokens and looked up in the dictionary. For words not in the dictionary, the necessary lexical information is predicted, including pronunciation and lexical stress [4], syllabification, possible parts-of-speech and syntactic roles. Then, sentences are part-of-speech tagged (allowing homograph disambiguation) and parsed and finally, the text is normalized (expansion of digits, abbreviations etc.). This concludes the text processing part of the system, whose output forms the input for the subsequent prosody prediction.

Prosody prediction consists of several completely data-driven modules that predict:

- the presence or absence of prosodic phrase breaks (chunk boundaries) [5];
- the presence or absence of pauses [5] [6];
- the length of previously predicted pauses [6];
- the accent property of each word: deaccented, accented or highly accented [6];
- continuous speech effects and speaker-specific pronunciations [7];
- the duration of each phone [6];
- the F0 contour and offset of each word [8].

2.2. Back-end

The "plural unit selection and fusion method" [9] is used to generate synthetic speech from the phone sequence, predicted prosody and some linguistic information given by the front-end. This method is a unit selection based method, in which speech units are selected from a large speech unit database using the cost function, but it is different from the conventional unit selection method. In the conventional unit selection method, a single speech unit is selected for each segment, and the selected units are concatenated with or without modifying the prosody of the units to generate synthetic speech. On the other hand, in the "plural unit selection and fusion method", several speech units are selected for each segment and are fused to generate a new speech unit for the segment. After modifying the prosody of these newly

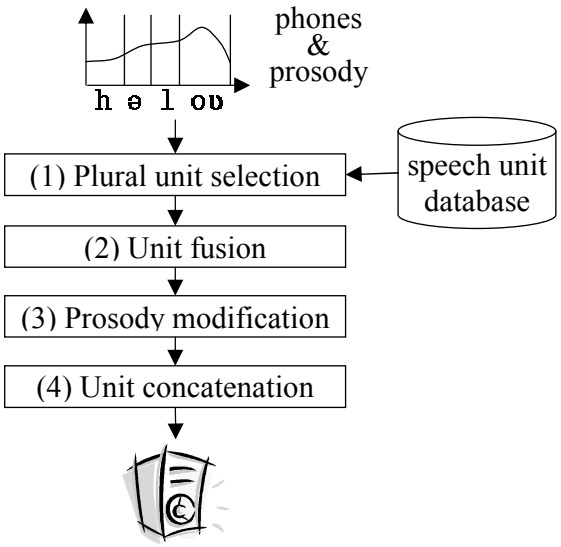

Figure 1: Flow of plural unit selection and fusion

generated speech units, they are concatenated to generate synthetic speech. The processing flow is depicted in Figure 1.

In the unit selection process, the selection of several speech units for each segment is performed in two steps. First, as in the case of the conventional unit selection method, the optimum speech unit sequence, which contains a single speech unit for each segment and has the smallest total cost, is selected using the DP (dynamic programming) algorithm. Next, based on the optimum speech units, N speech units are selected for each segment. In this selection, the target cost plus the concatenation costs with the previous and following optimum speech units is evaluated for each speech unit candidate for the segment except the optimum speech unit, and the (N-1) units with the (N-1) least costs are selected in addition to the optimum speech unit, from the speech unit candidates for the segment. In the cost function used in this process, F0, duration, phonetic context, etc. are considered as the target cost, and spectrum, etc. are considered as the concatenation cost.

In the unit fusion process, the selected N speech units for a segment are fused into one speech unit. Figure 2 shows how the units are fused in the case of N = 3 as an example. This is performed in the following three steps:

1. Each selected speech unit is decomposed into pitch cycle waveforms using a Hanning window with twice the length of each pitch period.

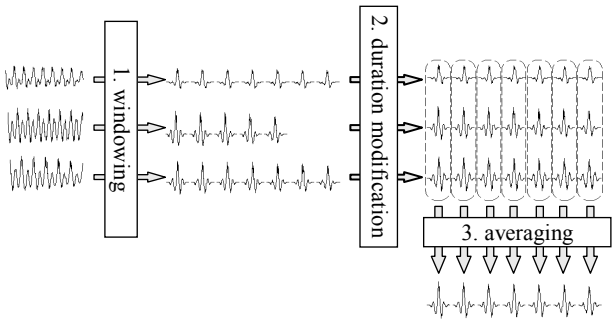

Figure 2: Unit fusion process.

2. For each speech unit, the number of pitch cycle waveforms is adjusted to that of pitch marks of the target, by duplication or elimination.
3. Pitch-cycle waveforms are averaged in the time domain.

In step 3, averaging is performed in each of 4 sub-bands. First, each pitch-cycle waveform is decomposed into sub-bands. Then, in each sub-band, the phase of each sub-band pitch-cycle waveform is aligned to match the one of the optimum speech unit, and then the aligned sub-band pitch-cycle waveforms are averaged. The phase alignment is realized by shifting the pitch-cycle waveform along the time direction so that the cross-correlation with the pitch-cycle waveform of the optimum speech-unit is maximized. Finally, the averaged sub-band pitch-cycle waveforms are summed to construct a pitch-cycle waveform of the fused pitch-cycle waveform. Actually, the explanation above is for a segment in a voiced phone. For a segment in an unvoiced phone, just the amplitude of the optimum speech unit is modified so that it should have the average power of the selected speech units for the segment.

The fused pitch-cycle waveforms generated above are then overlapped and added according to the pitch marks calculated from the target prosody to generate synthetic speech.

Although the conventional unit selection method can achieve high quality synthetic speech especially when the speech-unit database is very large, it tends to have some unstable parts in voice quality and discontinuities at unit boundaries. This is because suitable speech units for the target are not necessarily selected for all the segments, and, as a result, the voice quality can vary segment by segment in a sentence. Our method can improve these problems by fusing the speech units. The averaging operation in the unit fusion process reduces the voice quality fluctuation through a sentence and is expected to reduce the average distortion in voice quality [9]. An expected side-effect that the spectrum is smeared by averaging waveforms is limited since the averaging is performed in each sub-band. In addition, such smearing effect can also be compensated for to some extent by applying the formant emphasis filter to the fused pitch-cycle waveforms. As a result, our method can generate more stable and continuous synthetic speech than the conventional unit selection method while keeping the naturalness of the conventional one, thereby achieving very good intelligibility.

3. Voice corpus preparation

The following sections describe some automatic and manual steps that were applied to the voice corpus before the system was trained on it.

3.1. Power normalization

We found that the speech waveforms provided by the organizers had a problem in that their volume level fluctuated considerably among different waveform files. The waveforms seemed to have undergone some kind of power normalization, in which the waveform power was normalized so that the maximum amplitude should be approximately the same among the waveform files. The problem was probably a side-effect of the power normalization, because such a normalization would compress the amplitude too much for a waveform file with very large amplitude fluctuation, while it would allow a file with steady amplitude to be relatively loud.

Since this problem was assumed to have critically bad effects on our back-end, we applied a different power normalization to the provided waveforms. In this normalization, the amplitude of the waveform was modified linearly for each waveform file, so that the RMS (root mean square) of the amplitudes within the vowel sections in the file should be approximately the same among the waveform files. By this normalization, the volume levels of the waveforms got more consistent among the waveform files, as verified by listening to dozens of the files.

3.2. Automatic annotation

The Blizzard data came with automatically derived Festival utterance structures. However, as these annotations use different conventions than our system (e.g. a slightly more coarse-grained phone set and a different part-of-speech tag set), they are not straightforward to use for training our system. We therefore decided to create our own annotation (see below for a preliminary comparison of the two annotations). The annotation needed consists of part-of-speech and dependency parses, narrowly transcribed time-aligned phone sequences (including pauses), syllabification and position of lexical stress, prosodic chunk boundaries, sentence accents and pitch contours. To create it, we first extracted the pitch contour of each sentence using *get_f0* from the ESPS/waves toolkit [9]. We then tagged, parsed and normalized the corpus using our TTS system's text analyzer and text normalizer. We then looked up the canonical pronunciation of each word in the dictionary (see also Section 3.1.3). For the special case of acronyms (e.g. "AEC") we added a letter-by-letter pronunciation and, if the acronym contained at least one vowel, also a "normal word" pronunciation.

From these canonical pronunciations, we generated additional pronunciation variants by optionally applying phoneme deletion, insertion or substitution rules that were derived from the differences between canonical pronunciations and manually corrected narrow transcriptions for one of our in-house American voice databases. Each phoneme modification rule has an associated probability and the number of pronunciation variants to be generated can be controlled by using only rules with a probability above a certain threshold.

After pronunciation variant generation, an improved version of the Aligner [11] (based on HTK [12]) was used to decide between the variants and to time-align the resulting phone sequence, allowing for optional pauses after each word. As each pronunciation variant has an associated syllabification and lexical stress position, the appropriate information can be added for whatever variant has been chosen.

The Aligner does a so-called flat start, i.e. instead of using a pre-trained speaker-independent model it bootstraps a speaker-specific model from the data to be aligned. As the rules of the Blizzard Challenge for the two subset voices B and C do not allow participants to use any information from the full corpus outside the subset, we had to flat-start the Aligner for each of the three voices independently. As flat-starting on less data generally results in slightly worse alignments, we can expect the alignment for the B and C voices to be worse than that for the A voice.

After automatic alignment, we used a general tool, called the Prosodizer [13], to automatically derive ToBI annotation [14] and then mapped that annotation to the more coarse-grained markup that our system's prosodic chunker and sentence accent module need for training. The Prosodizer uses hand-written, language-specific but speaker-independent rules based on punctuation, the part-of-speech labels, the F0 contour and the phone alignment to derive ToBI labels. The output of the Aligner is thus one of the inputs to the Prosodizer. We therefore also had to apply the Prosodizer three times, for each of voice A, B and C, each time using the corresponding flat-started Aligner output as one of the inputs.

3.2.1. Comparison to pauses in the provided annotation

As there is no manually corrected annotation of the voice database, it is impossible to simply measure whether our automatic annotation is better than the one provided by the organizers. However, we conducted a preliminary investigation into the annotation of pauses. Table 1 shows the raw numbers of pauses in both annotations of the complete database (6,579 utterances).

Sentence internal pauses	organizers' annotation	our annotation
<100 ms	648	1372
>=100 ms & <200 ms	1035	1323
>200 ms	1541	2442
Total	3224	5137

Table 1: Number of sentence internal pauses in the annotation provided by the organizers and automatically derived by us.

As can be seen, our annotation contains 1913 pauses more than the annotation provided by the organizers, which is almost 60% more. In order to estimate the accuracy of these additional pauses we manually checked 20 randomly selected sentences which included pauses in our but not in the organizers' annotation. Of the 24 pauses in these 20 sentences, 21 were definitely correct, and 3 could be regarded as debatable. None was clearly wrong. In the organizers' annotation, these pauses were subsumed under word-final or word-initial phones. The missing pauses were not only short ones: sometimes pauses longer than 300ms were missing. In many of these missed pauses the pause included some sort of noise like a little smack or a breath noise. We conclude that at least in terms of pauses, it was worth producing our own annotation. It would be interesting to also compare other aspects of the annotation.

3.3. Unknown words

Our in-house American English pronunciation dictionary is quite extensive. For the Blizzard Challenge, we supplemented it with CMUDICT [15] converted to our phone set. We then checked all the words in the Blizzard corpus against this combined dictionary and identified all unknown words. It turned out that many of these were rightly absent from both American English pronunciation dictionaries.

7 of the unknown words were Spanish, stemming from three completely Spanish sentences in the voice corpus, which we excluded from our process. 60 were acronyms, for which no manual work is required (see Section 3.2)

Many unknown words were Japanese, mostly names but also common nouns such as “onsen” (hot spring). In contrast to probably the majority of American speakers, the Blizzard speaker does not pronounce these words in a very Americanized way but rather uses something close to their Japanese pronunciation. The unknown Japanese words were therefore manually transcribed by a native speaker of Japanese.
Of the remaining words, many were foreign names, or derivations thereof. More than half could be confidently transcribed by a native English speaker without using the audio files, while the other half were so obscure that even that person had to listen to the audio files and use whatever pronunciation the speaker used.
All of the manually transcribed words were added to the combined dictionary before the (final run of the) automatic annotation process described in the previous section started. Therefore, all words were known during that process and the normal letter-to-sound rules were not used.
As allowed by the Blizzard rules, we also manually transcribed the unknown words in the test sentences and added them to the dictionary before synthesizing these sentences.

3.4. Manual corrections

It would have been possible to train a complete system on the annotations derived by the steps described in Section 3.2. However, there was time left and it is likely that manual correction of the biggest annotation errors improves performance of the resulting system. We therefore checked and, if necessary, corrected

- the normalization of all digits in the input text, and
- periods not at the end of the input text remaining after text normalization (most were sentence end periods from those BTEC texts containing more than one sentence but a few were missed abbreviations)

The method of F0 prediction of our TTS system uses a separate set of codebook entries to model the rising pitch that typically marks the end of yes/no-questions in English [16]. The codebook entries are trained on words annotated with the H-H% ToBI label. As yes/no-questions are relatively rare, it is especially important that their annotation is correct. Therefore, all sentences to which the Prosodizer had assigned this label were checked (to find false positives), as were all sentences ending in a question mark but not starting with a question word (to find false negatives).
The manual work on the corpus also uncovered at least 14 cases in which the speaker deviated from the script (e.g. saying “located” instead of “based”) and about 16 where a final plosive had been (partially) cut off from the recording. Most of these cases were corrected as well (by modifying the script or the transcription).

4. Training and parameter optimization

Once the data has been annotated, training new prosody models and creating a new unit database is a completely automatic process. Nearly all scripts for training speaker-specific modules are integrated into one Makefile, which allows for easy partial retraining, if necessary. The only exception (due to a lack of time) is the module for training the prediction of continuous speech effects. It has therefore not been retrained for Blizzard; rather the model derived from one of our in-house American English voice databases is used.

4.1. Optimization of unit-fusion parameters

The number of speech-units to be fused, N, can be changed to control the trade-off between stability and segmental naturalness of the synthetic speech. Increasing N can enhance the stability while it may sacrifice the segmental naturalness. In addition, the strength of the formant emphasis filter that is applied to the fused pitch-cycle waveforms affects the segmental naturalness. The appropriate strength would be related to N since it would depend on how much the fused pitch-cycle waveforms are smeared by averaging.

Therefore, the combination of these two parameters was optimized for the Blizzard narrator using voice B. Out of the several combinations of the two parameter values, the best combination was selected by an informal listening test in which MOS was evaluated for every combination. The combination of N=10 and a strong setting of the formant emphasis filter was selected as the best.

5. The three voices

As described in the introduction, the 2007 Blizzard Challenge consisted of three subtasks, one involving the full corpus and two involving subsets of it, one given and one selected by each participant. We submitted voices for all three conditions.

5.1. Voice A

For voice A, we annotated and trained on the complete corpus. The whole automatic process took about 4 days. None of the tools used for annotation and training is particularly optimized for speed, so it is likely that the process could be accelerated considerably if necessary. Transcription of unknown words and manual checking and correcting of automatic annotations was done by several different people at different times and is therefore difficult to time.
Synthesizing the 400 tests sentences sent by the organizers with the TTS system based on voice A took less than half an hour.

5.2. Voice B

The Arctic corpus was designed to be phonetically balanced but it does lack coverage of certain other aspects important for training our system. It contains only a single question, and no yes/no-question, so the separate codebook for modelling these could not be trained. We therefore decided to copy these codebook entries from a codebook trained on the question corpus used in [16] (as allowed by the rules).
The Arctic database also contains very few colons and semi-colons so our data-driven chunker and pause prediction cannot be trained properly for these cases. We therefore added a manual rule to enforce a prosodic break and a pause at the punctuation marks. Note that none of these manual interventions were necessary for voice A, where the appropriate pitch contours and chunk/pause rules could simply be derived automatically from the data.

5.3. Voice C

For voice C, we randomly selected utterances from the full corpus until the allowed duration was reached. This resulted in

687 utterances. We did try an alternative (a phonologically balanced selection) but a preference test revealed that the voice based on the random selection was preferred; so we decided to use the latter as our Blizzard submission for voice C. See [17] for a detailed description and analysis of these experiments.

6. Results and discussion

Table 2 shows our Word Error Rate (WER), based on the 16 semantically unpredictable (SUS) sentences used in the official evaluation, and our mean opinion score (MOS), based on the 34 news and conversation utterances, for each of the three voices. Results for other participants can be found in [1]. While the organizers provide a ranking, they also stress that many differences are not statistically significant. Based on their table of statistical significance between systems, we grouped systems as follows: Let *i*=1. Take the highest-ranked still ungrouped system and all other ungrouped ones whose MOS resp. WER is not significantly different and put them into group *i*. Increment *i* and repeat until all systems are grouped.[2] Table 2 also shows which group our system is in for the two metrics and three voices.

As can be seen, our WERs are always among the best, in fact, our WER for voice C is the lowest of all systems and for voice B, it is a shared lowest (although not statistically significant from some other systems).

On a relative basis, the MOS are slightly better for the subset voices (B, C) than for the full voice (A).

These results confirm the effect of our normal focus on good intelligibility with a small footprint (which typically is much smaller than for the Blizzard Challenge, where no maximum was imposed).

They also confirm that our simple approach of random sentence selection for voice C has worked.

		A	B	C
Intelli-gibility	WER	0.26	0.25	0.34
	in group	1st of 4	1st of 4	1st of 3
Natural-ness	MOS	3.0	2.9	2.9
	in group	3rd of 5	2nd of 6	2nd of 6

Table 2: Intelligibility and Naturalness results of our system, based on all listeners.

7. Conclusions and future work

We have described the system with which we took part in the Blizzard Challenge for the first time. Instead of using the provided annotation, we created our own from scratch, including the generation of narrowly transcribed pronunciation variants and the automatic prediction of ToBI labels. A preliminary comparison of the pauses in our annotations with the provided annotations shows that our method correctly finds many more pauses. Our back-end uses the "plural unit selection and fusion method". For the Blizzard Challenge, 10 units were fused into one for each voiced unit.

Results show that our system achieves excellent intelligibility, in particular for the smaller voices.

The obvious next steps are a more detailed error analysis and research into improving the naturalness without sacrificing intelligibility. In particular, we think that our system might not fully exploit the possibilities offered by a relatively large voice database combined with an unlimited footptrint (an unusual luxury for us).

[2] The resulting group sizes are:

for WER:	voice A:	12 - 2 - 1 - 1
	voice B:	10 - 3 - 2 – 1
	voice C:	9 - 1 - 1
for MOS:	voice A:	2 - 2 - 4 - 3 - 3
	voice B:	5 - 3 - 4 - 1 - 1 - 2
	voice C:	4 - 3 - 1 - 1 - 1 - 1

8. Acknowledgements

We would like to thank Zoe Handley for handling the dictionary work, Jessica Müller for working on the text normalization and all our colleagues who took part in the evaluations. We are also grateful to ATR for making the voice corpus available for the Blizzard Challenge.

9. References

[1] Fraser, M., and King, S., "The Blizzard Challenge 2007", *Proc. Blizzard Workshop (in Proc. SSW6)*, August 2007, Bonn, Germany.

[2] Kominek, J. and Black, A. W., *"CMU Arctic Databases for Speech Synthesis"*, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2003.

[3] Ni, J., Hirai, T., Kawai, H., Toda, T., Tokuda, K., Tsuzaki, M., Sakai, S., Maia, R., and Nakamura, S., "ATRECSS - ATR English speech corpus for speech synthesis", *Proc. Blizzard Workshop (in Proc. SSW6)*, August 2007, Bonn, Germany.

[4] Webster, G., "Improving Letter-to-Pronunciation Accuracy with Automatic Morphologically-Based Stress Prediction", *Proceedings of InterSpeech*, 2004.

[5] Burrows, T., Jackson, P., Knill, K., and Sityaev, D., "Combining Models of Prosodic Phrasing and Pausing", *Proceedings of Interspeech, 9th International Conference on Speech Communication and Technology*, p 1829–1832, 2005.

[6] Suh, C., Kagoshima, T., Morita, M., Seto, S., and Akamine, M., "TOSHIBA ENGLISH TEXT-TO-SPEECH SYNTHESIZER (TESS)", *6th European Conference on Speech Communication and Technology (EUROSPEECH'99)*, 1999.

[7] Webster, G., Burrows, T., and Knill, K., "A Comparison of Methods for Speaker-Dependent Pronunciation Tuning for Text-to-Speech Synthesis", *Proceedings of InterSpeech*, 2005.

[8] Kagoshima, T., Morita, M., Seto, S., and Akamine, M., "An F0 Contour Control Model for Totally Speaker Driven Text to Speech System", *Proceedings of ICSLP'98*, p 1975-1978, 1998.

[9] Mizutani, T., and Kagoshima, T., "Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method", *IEICE Trans. Vol. E88-D*, no.11, p 2565-2572, 2005.

[10] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", In W.B. Kleijn and K. K. Paliwal (Eds.),

Speech Coding and Synthesis, Amsterdam, Netherlands, Elsevier Science, 495–518, 1995.

[11] Talkin, D., and Wightman, C. W. "The Aligner: Text to speech alignment using Markov models and a pronunciation dictionary", *Proceedings of 2nd ESCA/IEEEWorkshop on Speech Synthesis*, p 89–92, 1994.

[12] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. , Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P. "*The HTK Book*", (for HTK Version 3.4), Cambridge, United Kingdom, 2006.

[13] Braunschweiler, N., "The Prosodizer – Automatic Prosodic Annotations of Speech Synthesis Databases", *Proceedings of Speech Prosody 2006,* PS5-27-76, 2006.

[14] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. "ToBI: A Standard for Labeling English Prosody", *Proceedings of the International Conference on Spoken Language Systems*, p. 867–870, 1992.

[15] CMU, "*Carnegie Mellon Pronouncing Dictionary*", http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 1998.

[16] Sityaev, D., Burrows, T., Jackson, P., and Knill, K., "Analysis and Modelling of Question Intonation in American English", *Proceedings of Speech Prosody*, 2006.

[17] Lambert, T., Braunschweiler, N., and Buchholz, S., "How (Not) to Select Your Voice Corpus: Random Selection vs. Phonologically Balanced", *Proceedings of the 6th ISCA Speech Synthesis Workshop (SSW6)*, 2007.