# How (Not) to Select Your Voice Corpus: Random Selection vs. Phonologically Balanced

*Tanya Lambert* [§], *Norbert Braunschweiler* [‡], *Sabine Buchholz* [‡]

‡ Speech Technology Group, Cambridge Research Laboratory,
Toshiba Research Europe Ltd., Cambridge, United Kingdom

tlambert@freeola.net, {norbert.braunschweiler, sabine.buchholz}@crl.toshiba.co.uk

## Abstract

This paper compares the effect of two different voice corpus selection methods on the overall quality of unit selection-based text-to-speech (TTS) voices resulting from training on these corpora. The first selection method aims to maximize the coverage of stressed as well as unstressed diphones (phonologically balanced: *Phonbal*) while the second method simply selects sentences at random (*Random*). We show that, as expected, the *Phonbal* method results in better phonetic and phonological coverage for the training as well as unseen test sentences. However, we also provide evidence from an objective evaluation and a subjective listening test that the *Random* method results in an overall better voice quality when only automatic corpus annotation tools (such as forced alignment) are used, and potentially even with manual annotation. This result has general implications for the fast creation of TTS voices.

## 1. Introduction

For corpus-based text-to-speech systems, the quality of the corpus is one of the important factors of the resulting TTS voice quality. Corpus quality in turn has several independent factors: the suitability of the voice talent, the quality of the recording, the quality of the annotation, and the choice of sentences to be recorded. This paper reports about experiments and analyses concerning this last factor. Traditionally, sentences have been chosen to maximize the diphone coverage [1, 2]. Recently, this approach has been extended to the coverage of diphones in stressed as well as unstressed positions, henceforth called "lexical diphones" [3, 4]. However, it is not clear whether this approach is optimal for all types of corpus-based TTS systems. This paper presents a case study aimed at answering the following question: what is the effect of different sentence selection methods on a halfphone-based unit-selection system with fully corpus-based prosodic components when only automatic corpus annotation is used? In particular, we compare two methods: one in which sentences are sampled at random from a much larger corpus [5] and another in which sentences are chosen in order to maximize the coverage of lexical diphones. Section 2 describes the background of this work and the two methods in detail. Section 3 compares the phonetic and phonological coverage of the two sub-corpora, while Section 4 compares the sub-corpora in terms of other aspects that are important for training a TTS voice, in particular phonetic alignment and prosody. Section 5 describes the listening tests that were conducted to compare the overall quality of the voices based on the two sub-corpora. Finally, Section 6 presents conclusions and plans for future research.

§ Affiliated to Toshiba when the work reported in this paper started.

## 2. Selection of sub-corpora

The experiments described in this paper took place in the context of the Blizzard Challenge 2007 [6]. Participants in this Challenge received the "ATR American English Speech Corpus for Speech Synthesis" [5], henceforth referred to as the *Full corpus*. It consists of utterance-length audio files totalling about 8 hours, corresponding text files and automatically created annotation. As the annotation supplied uses different conventions from what our system expects, we decided not to use that annotation but automatically created our own. See Section 4 for a summary of that method.

The *Full* corpus consists of sentences from three text genres: conversational (*BTEC*), news, and novels (*Arctic*). Upon receiving the *Full corpus*, participants had 4 weeks to create 3 TTS voices: one from the *Full corpus*, one from the *Arctic* sub-corpus, and one from a sub-corpus consisting of sentences that could be chosen freely from the *Full corpus* on condition that their total duration did not exceed the duration of the *Arctic* sub-corpus [7], which is 2,914 seconds (0.8 hours), and that the selection process does not rely on the audio files in any way. Our general system and results for the Blizzard Challenge 2007 are described in our Blizzard workshop paper [8], whereas the present paper focuses on the third voice condition only. The motivation for this third condition is to simulate the situation that one faces if one wants to record a new voice: given limited resources (e.g. budget, time) for recording, what is the best set of sentences one could record? The following two sections describe the two different corpus selection methods that we investigated: phonologically balanced versus random selection.

### 2.1. A phonologically rich sub-corpus

The phonologically balanced sub-corpus (*Phonbal*) was selected from the *Full* corpus using a greedy style set cover algorithm [3, 1]. This method focused on selecting lexical diphone types [3] from the *Full* corpus. A clear distinction is made between diphone types in stressed and unstressed lexical environments. For clarification, every phoneme in a phonetic string is assigned a lexical stress which it inherits from its parent syllable, e.g. /bs/ is a diphone type with no stress marking but /b0s1/ and /b1s0/ are lexical diphone types, where 0 and 1 indicate unstressed and primary stressed environments respectively.[1] The number of lexical diphone types in any text sample is much greater than the number of diphone types that are considered without stress markings. The text of the *Full* corpus was processed by Toshiba's TTS linguistic engine. Grapheme-

[1] Secondary, tertiary and/or emphatic stress could be considered in this way as well. However, as it was not used in the experiments described in this paper, it is ignored here.

to-phoneme conversion was performed and unstressed and primary stress assigned.

The creation of the phonologically rich sub-corpus initially focused on selecting all lexical diphone types present in the *Full* corpus. Based on the phonological transcription used here it was found that the *Full* corpus contained 368,039 lexical diphone tokens and 4,332 lexical diphone types. Lexical diphones also included silences (predicted from text-based features only). There were 631 lexical diphone types that appeared once in the text, and the most frequent lexical diphone type appeared over 8,500 times. The objective of the greedy style set cover algorithm was to capture the highest number of lexical diphone types within the smallest number of sentences. The phonologically rich sub-corpus generated in this way consisted of 1,133 sentences with speech duration of just over 6,000 seconds. As this is much more than the allowed 2,914 seconds, it had to be reduced in size.

The nature of the greedy-style algorithm is to rank sentences according to their phonological richness, where the lower ranking sentences cover only one unit of interest. In this selection, the lowest ranking 594 sentences (out of the 1,133) covered only one lexical diphone of interest. These 594 sentences were then reprocessed by excluding the primary stress information from lexical diphone combinations consisting only of consonants. The reason why the stress information was sacrificed in some consonant-consonant combinations is because past research has shown that any spectral discontinuities at concatenation points in the synthesis of CC (consonant-consonant) combinations are less likely to be detected aurally than in the synthesis of VC (vowel-consonant) combinations [9, 10].

As it was believed that different intonation types were necessary for the training of data used by the TTS system, some of the lexical diphone combinations were sacrificed at the cost of (i) intonationally rich phrases and (ii) consonant clusters preceeded and followed by a silence. With respect to (i) it was ensured that there was a sufficient coverage of interrogative sentences and multisyllabic words. With regards to (ii) consonant-vowel clusters preceeded by a phonetically marked silence (e.g. /#spɪ/, /#striː/) and vowel-consonant clusters followed by a silence (e.g. /ɪmd#/, /ɪkst#/ were added to the set. It was hoped that this inclusion would enable unit selection to choose phonetically and phonologically better suited consonants when synthesizing cluster combinations (i.e. to avoid the synthesis of e.g. /spl/ by combining /s/ and aspirated /pl/] or by combining /sp/ and clear /l/). In addition, it was hoped that this inclusion would offer better coverage with respect to falling or rising prosody depending on whether such clusters are preceeded or followed by a silence.

The phonologically rich corpus contained in the end a set of 728 sentences amounting to 2,906.25 seconds.

### 2.2. A randomly selected sub-corpus

The second sub-corpus was generated from the *Full* corpus by randomly selecting sentences until the maximum allowed duration was nearly reached. Then, a last sentence was selected that exactly filled the remaining duration. Therefore, the total speech duration for this *Random* sub-corpus equalled the *Arctic* speech database, i.e. 2914 seconds. The corpus consisted of 687 sentences.

Table 1 shows a comparison of the footprints of the *Full* corpus and its sub-corpora (*Arctic*, *Phonbal*, and *Random*) in terms of their duration, the number of sentences, words and words per sentence, the distribution of sentence lengths and the number of various punctuation characters.[2] The *Arctic* sub-corpus by design does not contain sentences of more than 15 words, which is why the average length and the distribution of sentence lengths are so different from the *Full* corpus. The lack of questions might be due to the nature of the text genre (novels).

Table 1: *Textual and duration characteristics of the Full corpus and its sub-corpora.*

| | *Full* | *Arctic* | *Phonbal* | *Random* |
|---|---|---|---|---|
| seconds | 28,591.5 | 2,914 | 2,906.25 | 2,914 |
| sentences | 6,579 | 1,032 | 728 | 687 |
| words | 79,182 | 9,196 | 8,156 | 8,094 |
| words/sent. | 12.0 | 8.9 | 11.2 | 11.8 |
| % sent. with | | | | |
| 1-9 words | 37.7 | 54.9 | 41.0 | 38.6 |
| 10-15 words | 27.6 | 45.1 | 18.6 | 26.9 |
| >15 words | 34.8 | - | 40.4 | 34.5 |
| '?' | 868 | 1 | 96 | 94 |
| '!' | 4 | - | - | 1 |
| ',' | 3,977 | 430 | 452 | 410 |
| ';' | 30 | 6 | 4 | 3 |
| ':' | 17 | - | - | - |

Table 2: *Unit type coverage in Full corpus and its sub-corpora.*

| Unit Types | *Full* | *Arctic* | *Phonbal* | *Random* |
|---|---|---|---|---|
| diph.(no stress) | 1607 | 1385 | 1510 | 1322 |
| lex. diphones | 4332 | 2716 | 3306 | 2735 |
| lex. triphones | 17032 | 7945 | 8716 | 8144 |
| sil_CV clusters | 104 | 42 | 46 | 43 |
| VC_sil clusters | 184 | 84 | 100 | 75 |

Among the two sub-corpora presented in this paper, *Random* is closer to the *Full* corpus than *Phonbal* in terms of average sentence length and the distribution of different sentence lengths. The greedy style set cover algorithm used to select the *Phonbal* seems to result in a greater number of short and long sentences being chosen, at the expense of the average-length ones. It remains to be investigated why this is the case. In terms of punctuation characters, *Phonbal* contains slightly more commas. This might be a side-effect of the presence of more long sentences.

## 3. Unit type coverage of the corpora

Table 2 shows the unit type coverage in the *Full* corpus and its sub-corpora. The distribution of unit types (diphones, lexical diphones, lexical triphones, silence_CV clusters and VC_silence clusters) is considerably smaller in the *Random* sub-corpus than in the *Phonbal* sub-corpus. In comparison with the *Arctic* speech database the random sub-corpus appears to have a better coverage of lexical diphone and lexical triphone types.

### 3.1. Coverage with respect to test sentences

400 test sentences provided by the Blizzard Challenge 2007 organizers were used here to objectively evaluate the phonological and phonetic coverage of the *Full* corpus and its sub-

[2] Counts for commas, semi-colons and colons are for sentence-internal ones only.

corpora. The test sentences comprised 100 sentences each from conversational (*conv*), *news* and *novel* text genres and 50 sentences each from modified rhyme tests (*mrt*) and semantically unpredictable sentences (*sus*). Figure 1 shows the coverage of diphone types (without stress consideration), lexical diphone and lexical triphone types in test sentences for each text genre. Occurrences of silence_CV clusters and VC_silence clusters in test sentences per given text genre are poor: less than 10 occurrences for silence_CV clusters types and less than 20 for VC_silence cluster types.

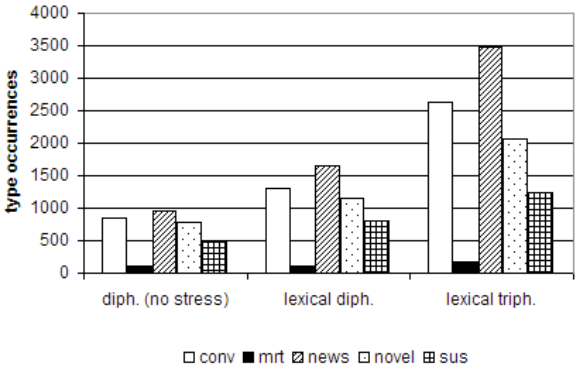


Figure 1: Distribution of diphone and triphone types in test sentences per text genre.

An analysis of unit coverage showed that neither the full corpus nor its sub-corpora contained all the lexical diphone types that were present in the 400 test sentences. Set cover mathematical operations (e.g. difference and intersection) were used on test sentences and (sub-)corpora to ascertain (i) which phonetic/phonological units were covered by both sets (Table 3), (ii) which units appeared in *Full* corpus/sub-corpora but were missing from the test sentences (Figures 2 and 3) and (iii) which units appeared in the test sentences but were missing from the *Full* corpus/sub-corpora(Figures 4 and 5).

Table 3: *Lexical diphone type coverage in the Full corpus and its sub-corpora for 400 test sentences.*

| TestSent | *Full* | *Arctic* | *Phonbal* | *Random* |
|---|---|---|---|---|
| conv | 1293 | 1203 | 1238 | 1212 |
| mrt | 113 | 111 | 111 | 108 |
| news | 1648 | 1518 | 1550 | 1534 |
| novel | 1147 | 1105 | 1099 | 1069 |
| sus | 799 | 741 | 760 | 739 |

With regard to the test sentences from novels, the *Arctic* sub-corpus appears to have better coverage than the *Phonbal* and *Random* sub-corpus. The *Phonbal* sub-corpus in comparison with the *Arctic* sub-corpus has better coverage of lexical diphone types with respect to test sentences for three text genres (for *mrt* there is a tie). In comparison with the *Random* sub-corpus, the *Phonbal* sub-corpus appears to have better lexical diphone type coverage for all five text genres.

Figures 2 and 3 show the number of diphone types that exist in the *Full* corpus and its sub-corpora but do not appear in the test sentences. For new and unpredictable test sentences, this figure indicates the phonetic and phonological richness of the given sub-corpus in relation to each text genre.

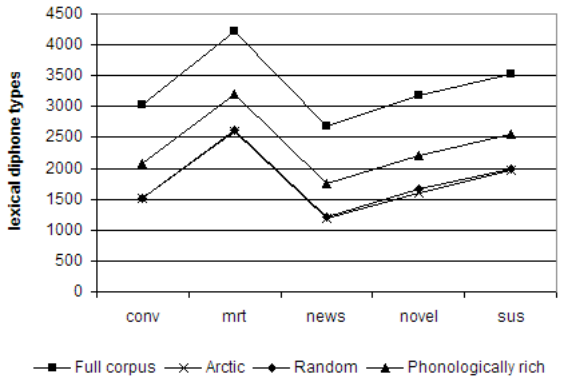


Figure 2: Lexical diphone types that appear in each (full/sub-) corpus but are missing from the test sentences.

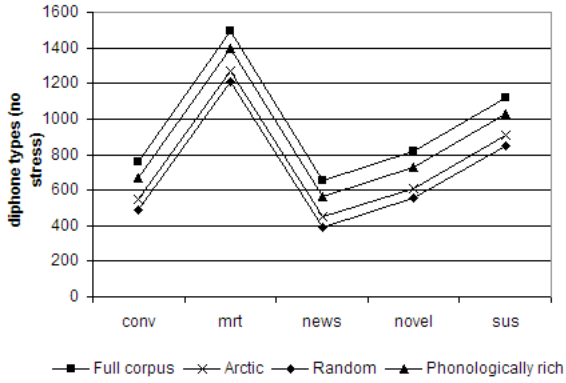


Figure 3: Diphone types that appear in each (full/sub-) corpus but are missing from the test sentences.

## 4. Objective evaluation

The previous section showed that the diphone and lexical diphone coverage of *Phonbal* is indeed better than that of *Random.* However, the corpus is used not only to derive the half-phones used by the TTS system but also to train its prosodic modules. The Toshiba TTS system contains a pipeline of modules that predict:

- the presence or absence of prosodic phrase breaks (chunk boundaries) [11];
- the presence or absence of pauses [11];
- the length of previously predicted pauses;
- the accent property of each word: deaccented, accented or highly accented;
- the duration of each phone;
- the pitch contour of each word.

The output of the pause, duration and pitch modules is used to restrict the unit selection (together with phonetic context and concatenation cost). If the selected units do not fulfil the target requirements of duration and pitch, they are modified accordingly. Therefore, the quality of the predicted prosody is an important factor in the overall voice quality.

All prosodic components are trained on the corpus. In the context of Blizzard, the corpus did not come with the necessary

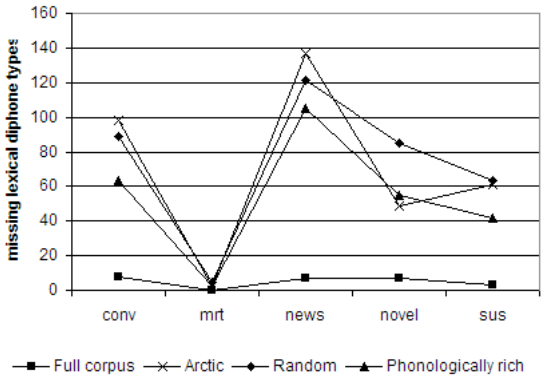


Figure 4: Lexical diphone types that are missing from each (full/sub-)corpus in relation to test sentences.

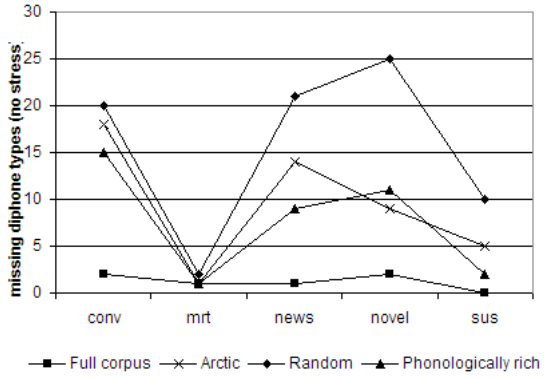


Figure 5: Diphone types that are missing from each (full/sub-) corpus in relation to test sentences.

annotations, and therefore had to be annotated automatically. First, the text was normalized (i.e. numbers, abbreviations etc. expanded). Next, likely phonetic transcriptions for each word were generated through a combination of lexicon-lookup and probabilistic post-lexical effects rules (to account for elision, assimilation etc.).[3] Then, an automatic phone aligner similar to the one described in [12] was used to perform forced phone alignment, choosing between potential pronunciation variants and allowing optional pauses at each word boundary.

As the Aligner did not use a pre-trained alignment model but rather performed a flat start [13] from the given corpus, the distribution of phones in a corpus potentially influences alignment quality. As no gold standard phonetic alignment is given for the *Full* corpus, we cannot directly measure the quality of the alignments in the two sub-corpora (*Phonbal* and *Random*). However, all other things being equal, flat starting on a larger corpus is very likely to result in better alignments than on a smaller one. It is therefore reasonable to assume that the alignments for the *Full* corpus are closer to the truth than those for the smaller sub-corpora. We therefore estimate the alignment quality of the sub-corpora by comparing them to the alignment of the *Full* corpus. We computed several metrics that have been suggested in the literature: overlap rate[4] [15], RMSE of phone boundaries,[5] and percentage of boundaries that are within certain tolerance margins of the "true" boundary. Table 4 shows the results.

Table 4: *Comparison of phone alignments in the Phonbal and Random sub-corpora against those in the Full corpus.*

| Metric | *Phonbal* | *Random* |
|---|---|---|
| Overlap Rate | 95.26 | 96.35 |
| RMSE of boundaries | 6.3 ms | 3.3 ms |
| boundaries within 5ms | 86.6 % | 91.8 % |
| boundaries within 10ms | 97.1 % | 99.1 % |
| boundaries within 20ms | 99.1 % | 99.9 % |

According to all metrics, the alignment of the *Random* subcorpus is slightly better than that of the *Phonbal* one. When comparing the overlap rate of individual phones, a similar picture emerges. The overlap rate of most phones is better for *Random* than for *Phonbal*; in particular, the overlap rate of *all* higher frequency phones (occurring more than 800 times in the two sub-corpora) is better for *Random*. Conversely, there are only 10 phones for which *Phonbal* has a better overlap rate, all of them of lower frequency. In *all* of these cases, *Phonbal* actually contains more instances of these phones than *Random*. In general, *Phonbal* contains more instances of rarer phones than *Random*, at the expense of more frequent phones. These figures suggest that a greater phonetic coverage of a sub-corpus has a detrimental effect on alignment accuracy. Interestingly, *Random* contains fewer sentence-internal pauses (and also fewer sentence-initial and sentence-final pauses because it generally contains fewer sentences than *Phonbal*) but the overlap rate of these pauses is much better than for *Phonbal* (97.72% vs. 86.95%; for sentence-initial/final pauses: 99.73% vs. 97.27%). More investigation is needed to explain this effect. Given that the phone duration and pause models are trained using the Aligner output, we can hypothesize that training on the *Random* corpus would result in slightly better pause and duration models. In addition, units derived from the *Random* corpus should generally have better boundaries, and might give rise to fewer bad joins during synthesis.

After forced alignment, the Prosodizer [16] is used to predict ToBI markup [14] based on the phone alignments, the previously predicted syntactic annotation and F0 contours extracted from the audio files using *get_f0* from the ESPS/waves toolkit [17]. The ToBI labels are then mapped to the more coarse-grained annotation on which the chunker and the accent module can be trained. The Prosodizer operates on the sentence level, which means that the *accuracy* of this annotation should be the same for both sub-corpora (contrary to what we saw for the Aligner). However, for training prosodic modules, it is also important that the training material contains a variety of prosodic contexts. Given that this concept is difficult to define,

[3] Words that did not occur in the pronunciation lexicon were mostly transcribed manually without reference to the audio. In some cases this was not possible because the transcriber did not know the word. Sentences containing these truly unknown words were excluded from the selection sub-corpora (in accordance with the Blizzard guidelines which forbid reference to the audio for corpus selection). For the *Full* corpus, the audio was consulted to transcribe those words. This means that neither the *Full* corpus nor the sub-corpora contain words whose pronunciations had to be derived by letter-to-sound rules. Therefore the quality of the transcriptions should be relatively high.

[4] The overlap rate "is the ratio between the number of frames that belong to that segment in both segmentations and the number of frames that belong to the segment in one segmentation".

[5] excluding boundaries where the sub-corpus and the *Full* corpus have non-identical phone labels

Table 5: *Precision and recall of pauses, prosodic chunk boundaries, and accented (acc) and highly accented (high) words predicted by the prosodic modules trained either on the Phonbal or on the Random sub-corpus against the automatic markup of 1000 sentences not belonging to either sub-corpus.*

| | | *Phonbal* | *Random* |
|---|---|---|---|
| Chunks | Precision | 58.9 | 56.3 |
| | Recall | 34.2 | 38.7 |
| Pauses | Precision | 63.1 | 63.4 |
| | Recall | 34.1 | 38.0 |
| acc | Precision | 69.7 | 69.5 |
| | Recall | 78.4 | 78.9 |
| high | Precision | 54.7 | 57.1 |
| | Recall | 38.6 | 41.1 |

we decided instead to measure the performance of the prosodic modules trained on both sub-corpora by comparing their predictions with the (automatic) annotations for 1000 sentences from the *Full* corpus which are neither in the *Random* nor in the *Phonbal* selection. Remember that the automatic annotation tools (Aligner and Prosodizer) heavily rely on the audio files, whereas the trained prosodic modules have to make their prediction from text-derived features only. It is therefore reasonable to assume that the more the predictions of a prosodic modules coincide with the automatic annotation, the better its performance.

Table 5 shows the precision and recall of (presence of) pauses, chunk boundaries, accented and highly accented words for the prosodic modules trained on each sub-corpus. For pauses and highly accented words, *Random* clearly has better performance: precision as well as recall are higher than for *Phonbal*. For chunks and normally accented words, *Random* has lower precision but higher recall than *Phonbal*. In these cases, it is unclear what the best balance between the two is. If one weighs both equally ($\beta = 1$) and computes the F-measure, *Random* has better performance (45.9 vs. 43.3 for chunks, 73.9 vs. 73.8 for accented). However, spurious chunk boundaries and accents are likely to have a bigger negative effect than missing ones, so $\beta = 1$ does probably not define the optimal trade-off. In any case, it is fair to say that some of the modules trained on the *Random* sub-corpus have a quantitatively better performance than those trained on *Phonbal*, whereas other modules are at least not clearly worse.

## 5. Subjective evaluation

The previous sections have shown which objective advantages and disadvantages the two sub-corpora have. However, objective metrics cannot yet replace subjective listening tests. We therefore conducted preference tests to determine which sub-corpus resulted in overall better voice quality. These tests included 53 sentences (25 relatively short declarative sentences, 11 longer sentences, 5 commands, 6 wh-questions, 6 yes/no-questions) which had been used in earlier listening tests independent of Blizzard. The sentences were synthesized with both systems and for each sentence, both samples were played one after the other. Subjects could listen to the stimuli repeatedly but were encouraged to give their answer after the first time. The order of sentences and the order of systems for each sentence were randomized for each listener. Subjects had to make a forced choice whether they preferred the first or the second sample.

Table 6: *Result of preference test comparing 53 test sentences synthesized with voice Phonbal or voice Random. Columns 2 and 3 show the number of times each subject preferred each voice.*

| Subject | *Phonbal* | *Random* |
|---|---|---|
| Non-American Listeners | | |
| 1 | 20 | 33 |
| 2 | 21 | 32 |
| 3 | 24 | 29 |
| 4 | 25 | 28 |
| All | 90 | 122 |
| American English Listeners | | |
| 1 | 21 | 32 |
| 2 | 21 | 32 |
| 3 | 16 | 37 |
| 4 | 23 | 30 |
| 5 | 25 | 28 |
| All | 106 | 159 |

In a preliminary test, 3 British and one German speech expert took part. A later, more formal test involved 5 American English speakers without specific speech technology knowledge. In the latter test, we also asked subjects to briefly write down theirs reason (if any) for each preference decision. Table 6 shows the quantitative results of both tests. Each of the 9 subjects preferred the Random over the Phonbal voice.

When comparing the preference scores for those sentences where either only *Phonbal* or only *Random* was missing a (non-lexical) diphone (6 and 9 sentences, respectively), we do observe that in general, the voice which has the diphone is preferred. However, as this effect concerns only a minority of sentences, and in any case *Phonbal* has only slightly fewer missing diphone tokens than *Random*, it does not change the overall picture.

In the future, we plan to analyze the comments by the American subjects in more detail, identify the points in the speech signals that caused them to prefer one version or the other and check whether we can trace them back to bad alignments or wrong prosody predictions.

## 6. Conclusions and future research

We have described the creation of two sub-corpora, a phonologically balanced (*Phonbal*) and a randomly selected one (*Random*), and have shown that listeners consistently prefer the TTS voice built with our system from the *Random* corpus. We have investigated the differences between the two sub-corpora and shown that although *Phonbal* has better diphone and lexical diphone coverage, the automatic phone alignment of the *Random* corpus is more accurate than that of the *Phonbal* one. In addition, the prosody predicted by the models trained on the *Random* corpus seems to be slightly better. We assume that these factors are at least part, if not all, of the explanation for the observed preference results.

The experiment described in this paper used a specific corpus, a specific (automatic) annotation method, and a specific TTS system. However, it is likely that other corpus-based unit-selection systems would also suffer quality losses when trained on worse alignments. This means that for the very fast creation of TTS voices, where one cannot manually correct the corpus

annotations, one should seriously consider how to select the set of sentences to be recorded.

In the future, we would like to explore the following questions in more details:

- Is the better prosody prediction performance only due to better automatic prosody annotation which is due to better phonetic alignment, or is the *Random* selection inherently better suited to train prosody models on, e.g. because its distribution of sentence lengths is not as skewed as the *Phonbal* one? This question can be answered by re-doing the automatic prosody annotation of the sub-corpora, but this time using the phone alignments of the *Full* corpus as input to the prosody annotation tool, thereby eliminating any difference in alignment quality, and then re-training the prosodic modules on the two sub-corpora. If the prosody predicted by the modules trained on the *Random* corpus is then still slightly better, the difference has to be inherent to the selection method. This would mean that a *Random* selection has advantages even when manual annotation is used, as long as the TTS prosody is trained on the corpus and not rule-based.
- What exactly is the relation between phone frequency and alignment accuracy?
- Why does the *Random* corpus have so much better pause alignment when it contains fewer pauses?
- Is it worth trying to construct some kind of prosodically balanced corpus to boost the performance of the trained prosody modules, or would that result in a similar detrimental effect on alignment accuracy?

## 7. Acknowledgement

We gratefully acknowledge the use of the "ATR American English Speech Corpus for Speech Synthesis", kindly supplied by ATR for the 2007 Blizzard Challenge. We would also like to thank all participants in the listening tests.

## 8. References

[1] François, H., and Boëffard, O., "Design of an Optimal Continuous Speech Database for Text-to-Speech Synthesis Considered as a Set Covering Problem", in *Proc. of Eurospeech'01*, Aalborg, Denmark, 2001, pp. 829–832.

[2] Beutnagel, M., Conkie, A., and Syrdall, A. K., "Diphone Synthesis Using Unit Selection", in *Proc. of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Blue Mountains, Australia, 1998, pp. 185–190.

[3] Lambert, T., "Databases for Concatenative Text-to-Speech Synthesis Systems - Unit Selection and Knowledge-Based Approach", Ph.D. dissertation, Univ. of East Anglia, 2005.

[4] Lambert, T., and Breen, A., "A Database Design for a TTS Synthesis System Using Lexical Diphones", *8th International Conference on Spoken Language Processing (ICSLP)*, Korea, 2004, pp. 1381–1384.

[5] *ATR American English Speech Corpus for Speech Synthesis*, Advanced Telecommunications Research Institute International (ATR), 2005–2007.

[6] Black, A., Tokuda, K., and King, S., "Blizzard Challenge 2007", in *Proc. of The 6th ISCA Speech Synthesis Workshop*, Bonn, Germany, 2007.

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

[8] Buchholz, S., Braunschweiler, N., Morita, M., and Webster, G., "The Toshiba entry for the Blizzard Challenge 2007", in *Proc. of The 6th ISCA Speech Synthesis Workshop*, Bonn, Germany, 2007.

[9] Syrdal, A. K., "Phonetic Effects on Listener Detection of Vowel Concatenation", in *Proc. of Eurospeech'01*, Aalborg, Denmark, 2001, pp. 979–982.

[10] Klabbers, E., and Veldhuis, R., "Reducing Audible Spectral Discontinuities", *IEEE Transactions on Speech and Audio Processing*, 9(1), pp. 39–51, 2001.

[11] Burrows, T., Jackson, P., Knill, K. and Sityaev, D., "Combining Models of Prosodic Phrasing and Pausing", in *Proc. of Interspeech*, 9th International Conference on Speech Communication and Technology, Lisboa, Portugal, 2005, pp. 1829–1832.

[12] Talkin, D., and Wightman, C. W., "The Aligner: Text to speech alignment using Markov models and a pronunciation dictionary", in *Proc. of 2nd ESCA/IEEE Workshop on Speech Synthesis*, Mohonk, New Paltz, NY, USA, 1994, pp. 89–92.

[13] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woddland, P., "The HTK Book", (for HTK Version 3.4), Cambridge, United Kingdom, 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", in *Proc. of the International Conference on Spoken Language Systems*, Banff, Canada, 1992, pp. 867–870.

[15] Paulo, S., and Oliveira, L. C., "Automatic Phone Alignment and its Confidence Measures", in *Proc. of Advances in Natural Language Processing*, 4th International Conference, EsTAL, Alicante, Spain, 2004, pp. 36–45.

[16] Braunschweiler, N., "The Prosodizer - Automatic Prosodic Annotations of Speech Synthesis Databases", in *Proc. of Speech Prosody*, 3rd International Conference, Dresden, Germany, 2006, PS5-27-76.

[17] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", in *Speech Coding and Synthesis*, W.B. Kleijn and K. K. Paliwal, Eds., Amsterdam, The Netherlands: Elsevier Science, pp. 495–518, 1995.