

Crafting Small Databases for Unit Selection TTS: Effects on Intelligibility

H. Timothy Bunnell

Center for Pediatric Auditory and Speech Sciences, Nemours Biomedical Research, USA

bunnell@asel.udel.edu

Abstract

When creating unit selection voices for personal use, e.g., for use in communication aids, it is often desirable to keep the speech database as small as possible. The present study examines the effects of database size and database content on the intelligibility of synthetic speech produced by the latest version of the ModelTalker TTS system. Intelligibility here is measured objectively with an open response SU sentence task. While previous work has examined similar questions, that work has typically been with an eye toward completeness of the database coverage and using tasks that assess perceptual quality, but not explicitly intelligibility.

Index Terms: speech synthesis, unit selection, database size, database content, intelligibility, personal synthetic voices

1. Introduction

Creation of personal unit selection synthetic voices can be viewed as an optimization problem in which synthetic speech quality—for the moment loosely defined—is the dependent measure to be optimized, and a potentially large number of factors relating to both the speech database content and the unit selection synthesis process must be balanced to achieve the optimal quality. Among the controllable factors related to database content are speech inventory size, inventory composition, and annotation accuracy. Note that what may be the most important determinants of quality in terms of database content—the talker's voice and speech characteristics—are not controllable factors for personal voices because the speech must be (or closely resemble) that of the person who wishes to create the voice.

The current study explores two of these factors, inventory size, and inventory content because they are of particular importance to individuals for whom recording a large inventory of speech is either impractical or impossible, but who may be able to record smaller inventories without difficulty. These would include patients with neurodegenerative diseases such as ALS and Parkinson's Disease, whose speech may already be mildly dysarthric, typically developing children who are helping to create a real child's voice for other children to use, or nearly anyone who is not a paid professional voice talent and wishes to develop a personal voice.

The tradeoffs among inventory size, inventory content, and concatenative synthetic speech quality have been widely recognized and discussed [1-5]. Most often, the primary goal in designing an inventory to record for unit selection synthesis is to achieve coverage of as large and representative a sample of units in the target language as possible. For recent research and commercial systems, this entails collecting a very large sample of texts that cover a variety of domains and that may additionally promote diversity in the speaking style of the talker. It is not unusual for such inventories to correspond to several hours of running speech.

While this approach demonstrably leads to very high quality and natural sounding unit selection voices, it is

impractical and in many cases truly impossible for individuals aiming to create a personal voice. This leads us to ask what is the smallest amount of speech that can be recorded to produce a unit selection voice of *acceptable* quality.

One approach to this is exemplified by the design of the ARCTIC inventories[4]. For these, a large corpus of text was analyzed to determine the number and frequency of units (diphones with stress-conditioned vowels) in the complete set. Subsets of the complete set were then selected with a greedy algorithm to provide complete coverage of the diphone units in the superset in as few sentences as possible. This greedy selection was run twice, providing two subsets and then the subsets were pruned to remove awkward or otherwise questionable sentences. The final result was a set of about 1100 sentences that varied in length up to 20 words.

For the ModelTalker project, we wanted to have a list of sentences with properties similar to those of the ARCTIC inventory, but with a few additional constraints. Most importantly, because the sentences might be recorded by children or adults with limited vocal capabilities, we wanted sentences that were shorter than typical ARCTIC sentences. Additionally, we wanted to design an inventory in which the sentences were ordered such that initial sentences were more important to the final quality the synthetic voice than were later sentences. This could effectively allow individuals who are recording their own speech for a personal synthetic voice to decide how they wish to weigh recording effort against final voice quality.

In the following the methods used for constructing an initial list of 3165 candidate sentences are described along with procedures used to select ordered recording inventories from this initial list, and an experiment is described where the intelligibility of these inventories is assessed with a semantically unpredictable sentence (SUS) task.

2. Method

2.1. Initial sentence selection

Texts from Project Gutenberg [6] were used to create an initial sentence list. Because children are among those who might be recording inventories for ModelTalker, the texts included a number of stories that might be read by or of interest to children. These were the *Velveteen Rabbit*, two of the *Wizard of Oz* stories, *Little Women*, and *White Fang*. A perl script was written to extract sentence-like fragments (hereafter referred to as sentences) from these texts and the sentences were passed through the ModelTalker TTS front end to identify and either fix or eliminate sentences that contained words or other tokens that the TTS system failed to pronounce correctly. From the resulting list, a set of about 4000 sentences was selected by randomly choosing about 1000 sentences from each of the four authors' texts (the *Velveteen Rabbit* provided only about 350 sentences).

All of the nearly 4000 sentences were then recorded by an adult male voice talent using a Sennheiser HD-410 head mounted microphone in a sound attenuated chamber. These

sentences were further screened to identify sentences that the voice talent found difficult, made mistakes on, or on which equipment difficulties led to distortions. In the end, a total of 3165 sentence were retained as part of the initial corpus. Hereafter, the list of 3165 sentences will be referred to as the full inventory and the associated recordings of those sentences as the full corpus.

2.2. Greedy Selection Algorithm

This full inventory was reordered in two different ways using a simple greedy algorithm. This algorithm starts with an empty "output" set of utterances and full "input" set of utterances. On each iteration, it adds to the output set the sentence from the input set that provides the largest number of missing phonetic "units", and removes that sentence from the input set. The algorithm always begins with units corresponding to single phones in the phone set used by the SRL TTS system. When enough sentences have been selected that there are no missing monophone units in the output set, the algorithm switches to longer units, either diphones, or triphones, and the algorithm is repeated for those units. If diphones are selected as the second stage, the algorithm switches to triphone units once all diphones that are present in the input set are included at least once in the output set. Thus, the algorithm always terminates with a triphone stage that is started either directly following the monophone stage, or after exhausting the set of unique diphones in the input set. In the sorting process, when there are ties, that is, more than one sentence that provides the same number of new units to the output set, the sentence that will contribute the smallest number of duplicate units is selected first.

2.3. Stimuli

2.3.1. Synthetic Voices

For this experiment, we ran the greedy selection algorithm both with and without the intermediate diphone selection stage. That is, in one case, the order of sentences in the output first achieved complete coverage of all monophone segments, then complete coverage of all diphone segments, and finally added triphone segments in order of decreasing number of new triphones per sentence. In the other case, the algorithm shifted from complete monophone coverage to immediately adding new utterances based on the number of new triphones they contributed. This provided two ordered lists of the full set, one in which the order favored achieving broad diphone coverage in as few sentences as possible, the other in which the order favored achieving broad triphone coverage in as few utterances as possible. Hereafter, these will be referred to as the DIFIRST and TRIFIRST sets respectively. Since both orderings were drawn with respect to the content of the 3165-sentence full inventory, there is no difference in the content of the total output set for either order. However, when small subsets of the complete set are drawn from the initial parts of the lists, there can be substantial differences in the speech material selected. For example if the first 200 sentences in each list are considered, there are 59 sentences in common and 341 unique sentences out of the total possible 400 sentences (i.e., 200 of each type). In other words, a bit less than 15% overlap in the sentences of the two lists. When the subsets comprise about half the total number of sentences in the full set, the percentage of shared sentences is about 46%. Clearly, selecting early elements to achieve rapid triphone coverage within the set versus attempting to achieve rapid diphone coverage results in substantially different sentences being

selected early on in the process. Table 1 shows the proportion of overlap between the DIFIRST and TRIFIRST sets for each of the list lengths employed in this experiment.

Table 1. Percentage of overlap (shared sentences) between lists selected to maximize breadth of diphone versus triphone coverage within the list.

List Size	Overlap (%)
200	14.75
400	17.75
800	36.44
1600	46.47

From these two orderings of the sentence set, subsets consisting of the first 200, 400, 800, and 1600 sentences were selected and used to construct unit selection voices for the ModelTalker TTS system. These 8 voices plus one voice constructed from the complete 3165 sentence parent inventory, and a set of 100 natural speech tokens recorded by the same talker who recorded the database for a total of 10 "voice" conditions.

It should be noted that one of the design goals of the ModelTalker voice construction process is that it operates without manual intervention. Thus, each voice of different size and content was created from the input sentences starting from a default segmentation, which is adapted during the voice building process. No attempt was made to locate or correct alignment or segmentation errors in the final unit selection database. To the extent that small data sets might cause the adaptive alignment strategy or subsequent pruning of problematic units to become less robust, that effect can contribute to the final intelligibility of the voice.

2.3.2. Sentences

Stimuli for the listening experiment consisted of a set of 100 Semantically Unpredictable (SU) sentences [7, 8] both naturally recorded and synthesized with each of the nine synthetic voices described above. The sentence texts were generated by randomly assigning words of appropriate parts of speech to locations within five sentence frames. The words for random assignment were predominantly single syllable words, but also contained a sampling of 2, 3, and 4-syllable words. The sentence frames varied between 6 and 8 words in length, and a constraint was applied to the sentence generation such that no sentence exceeded 10 syllables in length. The generation process also maintained balance among the sentence frames so that the total 100-sentence set contained exactly 20 sentences of each frame type. Example sentences of each type are given in Table 2.

Table 2. Example SU sentences

Frame	Example Text
1	The bed dwelt from the thin plane.
2	The camp saved the force that walked.
3	The good head poured the horse
4	Why does the night bless the wrong house?
5	Suspect the law or the cell.

Stimuli were blocked into 10 lists in a partial Latin squares design. Five lists contained stimuli of the DIFIRST type plus the full 3165-sentence inventory. The other five lists contained stimuli of the TRIFIRST type plus natural speech. Each list was constructed to contain 20 sentences (4 examples of each frame type) rendered by one voice. Table 3 illustrates how voice conditions were distributed over sentences for the DIFIRST stimuli to form 5 lists that covered all sentences

produced by all voices. A similar set of lists was constructed for TRIFIRST stimuli with the exception that natural speech tokens were substituted for “full” inventory synthetic speech stimuli.

Table 3. List construction for listening experiment.

Sentence Number	DFIRST List number				
	1	2	3	4	5
1-20	d200	d400	d800	d1600	full
21-40	d400	d800	d1600	full	d200
41-60	d800	d1600	full	d200	d400
61-80	d1600	full	d200	d400	d800
81-100	full	d200	d400	d800	d1600

2.4. Listeners

The listeners were 30 undergraduate students from the University of Delaware who volunteered to participate for extra course credit. All students reported having no history of hearing or speech disorders and that American English was their native language.

2.5. Procedure

When participants arrived in the testing area, they were first asked to read and sign an IRB-approved consent form, which described the purpose of the experiment. In the instructions, participants were told they would hear meaningless but grammatically correct sentences consisting of only real English words. They were told to guess at words even if they were very uncertain what they heard. If listeners were completely unable to understand or remember any particular item in a sentence, they were encouraged to enter XX instead of the unknown/forgotten word.

For the experiment, listeners were seated in a sound-dampened room in front of a laptop computer and wore headphones connected via a USB audio interface to the computer. On each trial, listeners heard a single presentation of an SU sentence and then typed the sentence into the text entry field of a small program designed for stimulus presentation. The program allowed listeners to edit their typed input if necessary and then press the RETURN key or click a Next button to proceed to the next trial. Listeners were not able to replay stimuli, but were given as much time as desired to enter their response. The experiment was thus self-paced.

After receiving instructions, participants were given a short (10 trial) practice session to become familiar with the task. Materials in the practice session were SU sentences recorded by a female talker. Neither the sentences nor the talker were used for the main part of the experiment. All items on the practice list were natural speech, not synthetic.

Listeners were then assigned to a pair of 100-sentence lists, one list of DIFIRST and one list of TRIFIRST stimuli. Half of the subjects started the experiment with a DIFIRST list, and the other half began with a TRIFIRST list. Listeners were offered a short break period after completing the first 100-sentence list and before beginning the second list. Participants typically completed the entire experiment including practice in 45 minutes or less.

2.6. Data Analysis

Data for each trial were analyzed as the word-level edit distance between the listener’s response and the stimulus (i.e., the total number of insertions, deletions, and substitutions needed to map the words typed in response onto those of the stimulus sentence). The scoring program took into account and

accepted as correct any homonyms of the stimulus words (e.g., plain and plane were treated as equivalent), and also accepted a small number of common and unambiguous misspellings (e.g., light and lite). No additional attempts were made to interpret or guess at the intent of listeners’ typed responses. Given the assumption that typos and idiosyncratic misspellings would be randomly distributed with respect to the conditions of interest (i.e., differences in synthetic voice quality), this seemed to be the most conservative approach.

For the preliminary data analyses presented here, effects and interactions related to the sentence frame type and the between-subjects factors of list and list presentation order will not be tested. This still permits a balanced within-subjects comparison of the effects of the inventory content (DIFIRST versus TRIFIRST) and size (200, 400, 800, or 1600 utterances). Additionally, it will be possible to report performance on both the full 3165-utterance database and with natural speech, but the experimental design does not permit directly testing those factors as part of a single analysis of variance.

3. Results

Overall, listeners responded with a mean edit distance of .89 edits per sentence. This corresponds approximately to a 13.27% word error rate. Edit distances ranged from a low of 0.18 edits per sentence for natural speech, to a high of 1.59 edits per sentence for the poorer 200-utterance voice, corresponding to word error rates of 2.66 and 24.11 percent respectively.

A repeated-measures ANOVA for the balanced factors of database content and size (summarized in Table 4) revealed significant main effects of design and size as well as a significant interaction between these two factors. The means underlying these significant effects are plotted in Figure 1 along with the mean edit distance for the full 3165-utterance database and natural speech.

Table 4. ANOVA summary for the results of Experiment I

Effect	df	SSQ	MSQ	F	<i>p</i>
content	1	1.034	1.034	8.86	0.0058
Err(sub:content)	29	3.836	0.117		
size	3	18.715	6.239	95.28	<0.001
Err(sub:size)	87	5.696	0.066		
content:size	3	1.512	0.504	10.13	<0.001
Err(sub:content:size)	87	4.330	0.050		

Figure 1 illustrates that mean edit distance decreases with increasing database size as one would expect. Overall, as indicated by the significant main effect of content, the DIFIRST databases resulted in somewhat more intelligible speech (lower edit distances) than TRIFIRST speech. And as indicated by the significant main effect of size, both database design conditions show the expected decrease in edit distance with increasing database size. However, for the TRIFIRST databases, mean edit distance decreases very consistently with increasing database size, whereas for DIFIRST databases edit distance decreases by only a small amount from the 200 to 400 utterance databases, then shows a large decrease from 400 to 800 utterance database, and then again less rapid decreases in edit distance for increasing sizes. These differences in the rate of decline in mean edit distance as a function of database size account for the significant interaction between database design and size.

Figure 1 also shows for reference the mean edit distances associated with the full 3165-utterance database and natural speech.

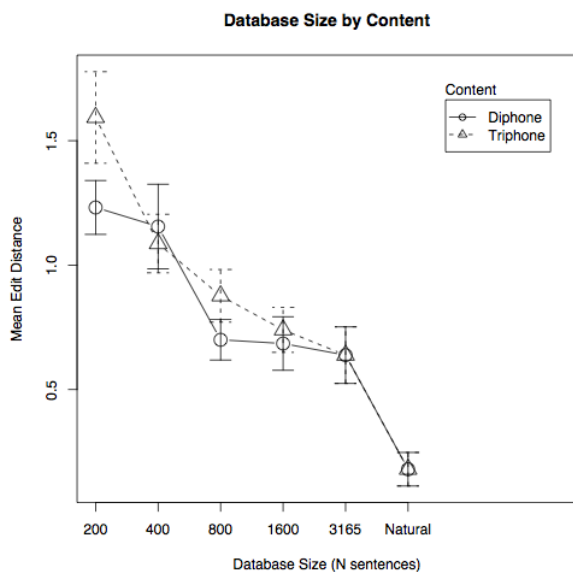


Figure 1: Mean edit distance between listener responses and intended utterances as a function of database size and content. Content was based on a DIFIRST strategy (circles) or TRIFIRST strategy (triangles). Error bars show the uncorrected 95% confidence intervals for each point.

4. Discussion

As expected, the intelligibility of these unit selection voices increased with increasing database size. On average, the edit distance was about 1.4 edits per sentence for SU sentences synthesized from 200-sentence databases. This corresponds to a word error rate of about 21.3%. The edit distance dropped to about .63 edits per sentence for the full 3165-sentence database, corresponding to a word error rate of roughly 9.4%.

It is difficult to say precisely what these error rates would correspond to in terms of meaningful sentence material. [7] reported a study in which they observed 85% intelligibility for synthesized high predictability SPIN sentences and estimated (based on extrapolation from sentence-level scoring of SU sentences) that the same TTS system produced 52% intelligibility with SU sentences. This relationship is quite close to that predicted between intelligibility for isolated phonetically balanced words and meaningful sentences (c.f., IEC standard 60849 common intelligibility scale). If that relationship holds for the current data, the better 200-sentence voice in the current experiment would provide better than 95% intelligibility for meaningful sentence material.

The present results are directly comparable to the results of a very similarly designed study reported in [8] that compared the intelligibility of five TTS systems using SU sentences. In that study, the best TTS system—an early 2000’s version of a large data-based commercial “voice”—had a word error rate of about 10.1% overall, and the worst system—the DECTalk Betty voice—a word error rate of about 34.6% overall. By contrast, the DIFIRST voices constructed from 800, 1600, and 3165 sentences in the present study achieved error rates of 10.4, 10.1, and 9.4 percent respectively. Thus, intelligibility of these voices approximates that of a commercial system that was based on several hours of recorded speech.

In the present results, the DIFIRST strategy for building database content achieved generally better results than did the TRIFIRST strategy. This led to the significant main effect of content with mean edit distances of .94 and 1.07 respectively for DIFIRST selected content versus TRIFIRST selected content. However, this overall effect was qualified by a significant interaction between content and database size, which is most evident in the case of the 400 sentence databases where listener responses to the sentences generated with the TRIFIRST database had slightly lower edit distances than responses to sentences generated from the DIFIRST database.

To better understand why the DIFIRST strategy only lead to poorer quality synthetic utterances with the 400-sentence database, the rate at which new units of various types entered the database was examined. This is illustrated in Figure 2, which plots the number of new units that are not already represented in the database that are added with each sentence up to 500 sentences. The type of units added, first monophones (M), then diphones (D), and finally triphones are used as the plotting symbols. Recall that for both DIFIRST and TRIFIRST strategies, monophones were first added until at least one instance of every unique monophone was present. This required 8 sentences. Then, for the DIFIRST strategy, sentences selected to add the largest number of new diphones we added until there was at least one instance of every unique diophone, at which point, the algorithm switched to selecting sentences for the number of new triphones.

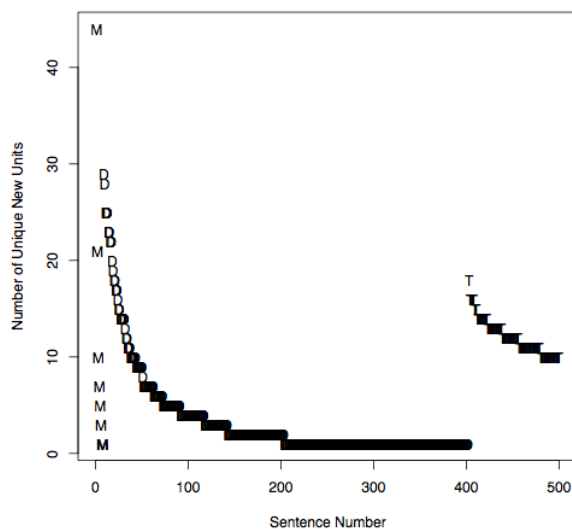


Figure 2. Number of new units added per sentence for the first 500 sentences. The type of unit is the parameter M(monophone), D(iphone), or T(riphone).

Of note in Figure 2 is the fact that from about sentence number 200 to sentence number 400, each new sentence added only one new diophone to the existing inventory. Thus, while the first 200 sentences formed a database comprising almost 5000 diphones, of which 1351 were unique, the second 200 sentences added only 203 more unique diphones for the 400-sentence database. Consequently, while almost doubling the database size, the 400-sentence database contained very little new phonetic material that was not already represented in the 200-sentence database.

Another facet of the database content in this study stemmed from our decision to define diophone and triphone

units in the simplest possible terms without regard to prosodic factors such as the stress of syllables from which the diphones were drawn or prosodic boundaries. For example, the ModelTalker prosodic model uses three levels of stress so that if all combinations of stress are possible for a given diphone, there would be 9 stress-conditioned variants for that diphone. In a similar vein, there are 5 levels of break index used in the ModelTalker prosodic model, and about 15 variants of pitch accents and tones. Used singly or in combination, these prosodic features can greatly extend the number of distinct diphone-sized units in the full database. There are just over 66,000 diphone tokens in the full database. Of these, only 1554 are unique when just phonetic identity is considered, but there are 13638 unique diphone variants when all prosodic features are considered as well. Exploration of whether adding diversity to the definition of a diphone by considering aspects of its prosodic structure will lead to improved inventories for constructing synthetic voices is currently ongoing in our laboratory.

5. Conclusions

When designing recording inventories for personal synthetic voices, it is typically desirable to keep the inventory size as small as possible while still achieving acceptable results. This is especially true for inventories recorded by patients with neurodegenerative diseases such as ALS for “voice banking.” In some cases, the onset of difficulty speaking or mild dysarthria is one of the initial symptoms leading to diagnosis of ALS. In many cases, patients may postpone voice banking until they notice some difficulty speaking. In either case, patients are unlikely to be able to record many hours of speech for a concatenative synthetic voice.

Even talkers whose speech is unaffected find it difficult to record large inventories of speech with the degree of consistency in speaking rate, voice quality, and pitch that is necessary to ensure a high quality voice. In that case as well, a small inventory that can be recorded in a relatively short period of time helps to minimize undesirable forms of variability.

From a user’s perspective, acceptability is defined among multiple dimensions. In our experience creating voices for ALS patients on the ModelTalker project, having a synthetic voice that sounds like the target talker is the primary goal in the sense that if this goal is not met, nothing else matters. Second to capturing talker identity, however, intelligibility seems to be most important. If one assumes that a minimally acceptable intelligibility score corresponds to a word error rate on SUS materials of not more than 20%, it appears that unit selection voices constructed with as few as 200 sentences can produce acceptable voices. For inventories larger than about 800 sentences (corresponding to only about 30 minutes of running speech) improvements in intelligibility appear to be very gradual. However, it should be noted that improvements in other aspects, notably prosody, may still be substantial.

6. Acknowledgements

This work was supported by grants from NIDCD and NIDRR and Nemours Biomedical Research. The author is grateful for the assistance of Jason Lilley and James Polikoff in collecting data.

7. References

1. Van Santen, J. and A. Buchsbaum, *Methods for optimal text selection*. Proceedings Eurospeech 97, 1997. 2: p. 553-556.
2. Black, A. and K. Lenzo, *Limited domain synthesis*. Proceedings of ICSLP 2000, 2000.
3. Black, A. and K. Lenzo, *Optimal data selection for unit selection synthesis*. Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW4), 2001: p. 63-67.
4. Kominek, J. and A. Black, *CMU ARCTIC databases for speech synthesis*. 2003: Pittsburgh, PA.
5. Francois, H. and O. Boeffard, *The Greedy Algorithm and its Application to the Construction of a Continuous Speech Database*. Proceedings of LREC, 2002: p. 1420-1426.
6. Hart, M.S. *Project Gutenberg Mission Statement*. 2007 [cited 2010 May 20]; Project Gutenberg home page]. Available from: <http://www.gutenberg.org>.
7. Benoît, C., M. Grice, and V. Hazan, *The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences*. Speech Communication, 1996. 18(4): p. 381-392.
8. Bunnell, H.T. and J. Lilley, *Analysis Methods for Assessing TTS Intelligibility*, in *SSW-6 The 6th ISCA Speech Synthesis Workshop*. 2007: Bonn, Germany. p. 374-379.