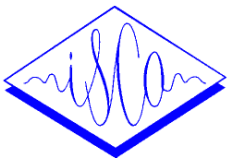

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

LIMITATIONS TO CONCATENATIVE SPEECH SYNTHESIS

Nick Campbell

ATR Spoken Langguage Translation Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
e-mail: toru-m@is.aist-nara.ac.jp, nick@slt.atr.co.jp

ABSTRACT

This paper discusses techniques for determining the linguistic needs for open-domain synthesis by concatenative methods, and reports on the design and evaluation of a tool for collecting and balancing a speech corpus automatically, in order to ensure optimal coverage of the sounds required for synthesis within a given task-domain. Synthetically-generated utterances are used to prompt speakers, and in-line acoustic analysis determines the prosodic as well as phonemic balance of the resulting speech during recording, re-prompting the speaker with textually modified versions if necessary, to elicit the desired articulation sequences. The closed-loop process, which incorporates human self-correction and evaluation, allows for more efficient collection of a balanced corpus for concatenative speech synthesis.

1. INTRODUCTION

With the advent of cheap memory and faster cpu-processing, concatenative methods for speech synthesis have now become standard. However, simply having a large source-unit database, and a fast machine with which to search it, does not guarantee high-quality results. The design of source texts for the recordings and the definition of the linguisic and para-linguistic situations that need to be covered by the resulting synthesis are crucial factors in determining the success of the synthesis.

This paper reports on the design and evaluation of a tool for collecting and balancing a speech corpus automatically in order to ensure optimal coverage of the sounds required for synthesis within a given task-domain, and discusses techniques for determining the linguistic needs for open-domain synthesis using concatenative methods.

For many task-specific speech synthesis purposes, a finite list of sentences can be recorded to provide balanced phonemic and prosodic coverage of the likely sound-sequences in the texts to be synthesised. However, the problem with recordings of such 'balanced' lists is that there is little control over the prosodic interpretation given to each utterance during the reading, since they are normally produced in isolation and out of context.

Previous work [1, 2, 3] described an apparatus to control the automatic construction of a source-unit corpus, when example speech waveforms are available, by using near-real-time acoustic analysis to determine the prosodic as well as the phonemic balance, re-prompting the speaker later in the recording sequence with textually modified (graphically annotated) versions of any mis-read prompt sentences in order to ensure the desired database content.

The current study details improvements to that design such that synthetically-generated prompt utterances can be effectively used in a similar way, taking account of the fact that human speakers naturally compensate for many of the unnatural artifacts arising from noise or inadequacy in the synthesised examples. By using the prediction modules of the synthesiser to generate and store the waveform and prosodic contours for each text sentence, the target utterances and their realised contours (pitch, power, and duration) can be compared. This closed-loop process, which includes human monitoring, allows for more efficient collection of a balanced speech corpus.

2. SPEECH SYNTHESIS TRENDS

There has been a shift of paradigm in approaches to speech synthesis. The history of this technology shows an evolution from compute-intensive limited-memory devices towards memory-intensive but light-load systems which make increasing use of natural speech sources for voice-creation. Figure 1 illustrates this trend. Whereas the majority of speech synthesisers in the nineteen-eighties relied on rule-based approaches, both for the prediction of an appropriate sound sequence and for the production of the speech waveforms, corpus-based developments throughout the nineties resulted in improved speech quality at the cost of increased memory usage. Improvements were gained when phone-based parametric prediction of waveform spectral and prosodic characteristics (e.g., MITalk [4], indicated by t1 in Fig.1) was substituted by diphone-based [5] and non-uniform [6] source units (t2 in Fig.1)), and similar improvements accompanied the move away from signal-processing for prosody modification towards raw-waveform concatenation methods [7, 8, 9] (t3 in Fig.1).

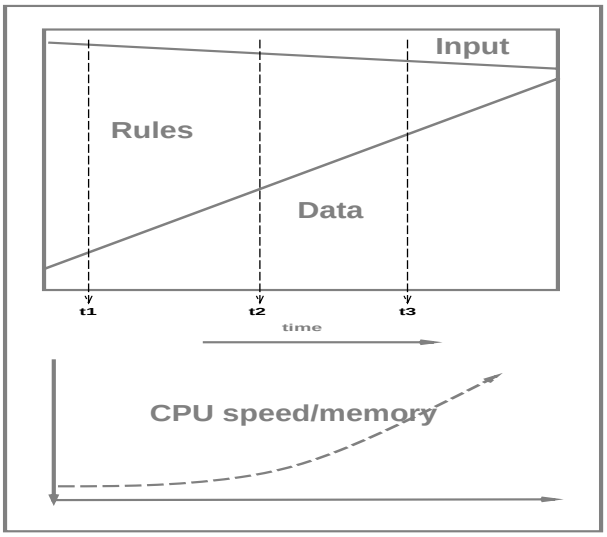

Figure 1: Trends in spech synthesis, showing that although computing power is increasing over time, improvements have been gained by replacing rule-based calculation with data-based knowledge.

It is of interest to note that the biggest increases in synthesis quality came not as a result of improved parametric modelling, or rules, but from the use of more informationally-rich source units i.e., data, which inherently include the supposedly redundant details that the parametric methods failed to predict. We can interpolate from the trends shown in the figure that future improvements will be gained from the inclusion of even more fine-detail from original speech, rather from improved modelling of the basic features.

With regard to future speech synthesis improvements (at t4, not shown but on the right in Fig.1), we can speculate that a further increase in the amount of available data will result in even less need for rule-based prediction, but with an increased demand for annotation of the input text (perhaps implemented in XML-markup as exemplified by SABLE [10]) in order to fully specify the output speech requirements. With an increase in the size of speech source-unit databases, we can foresee that the need for prediction of numerical values to specify the prosodic contours will be replaced by direct selection of appropriate units according to contextual factors (the same factors that are currently used for the prediction of the numeric specifications), but this requires a database which is rich in sources and contexts.

3. DATABASE PRODUCTION

There are two aspects to database production that need to be controlled for effective concatenative speech synthesis; the first is the content, and the second is the speaking-style. The content should be phonemically and prosodically balanced so that all variants of each sound can be represented. The speaking style should be appropriate for the task, as it is probably beyond the capability of current concatenative techniques to produce speech with a voice-quality that can match the full range of human expression (compare newsreading with after-dinner conversation for example).

With sufficient coverage in the source database, waveform manipulation can be minimised or eliminated for highest speech quality. Previous work has shown that controlled collection of speech data can be automated. Figure 2 shows the interface of an automatic speech collection device (DATR [1]) which prompts the reader with a series of sentences in both written and aural form. The text in this example is provided in three formats (Japanese Kanji characters, Japanese Kana phonetic alphabet, and Roman script with prosodic diacritics showing phrasing and accentuation) for an unambiguous interpretation, and a sample speech waveform is presented for each sentence so that timing and pitch characteristics can be reproduced. The speaker presses a button to start recording after listening to the sample and then compares the two waveforms aurally before saving the utterance and recording the next. The system compares waveforms by DTW to perform labelling of both prosodic and phonemic details while the reader is preparing the subsequent utterance.

Whereas the first version of this database recording software used human recordings of the sample utterances and was thus limited to reproducing available databases, albeit in new voices, a recent improvement makes use of synthesised 'originals' for the sample utterances, thus increasing flexibility considerably. Sentences can now be generated algorithmically, according to the current state of the database, to improve the balance and add new sound sequences. The synthesised prompt is a stylisation of the intended rendition, somewhat lacking in naturalness, but the version produced by the listener/speaker is richer in terms of timing and prosodic continuity, as well as in voice quality, being an interpretation of 'what the synthesiser was trying to say'. This improvement was not originally anticipated.

Separate work is being carried out to post-process the resulting speech database in order to distinguish differences in voice quality, resulting from different speaker states, which can influence voice-quality (such as tiredness, boredom or frustration), and a paper on that work will be presented at this conference[11].

4. DATABASE PRUNING

If the speech database contains two segments that are functionally identical (or perceptually equivalent) then only one token need be retained for synthesis, and duplicates should be pruned out for reasons of both

Figure 2: DATR's Recording Screen

elegance and efficiency. Previous work [14] has shown that objective acoustic measures of the distance between synthesised speech and its naturally-spoken original can correlate well with perceptual evaluations of the synthesis quality. The bi-spectrum [14] provides a good measure of distance between two signals that are phonemically equivalent. We attribute the efficiency of this measure to the fact that it incorporates phase information, which is particularly important for concatenative synthesis techniques.

In conjunction with the acoustic measure of similarity, we also employ a weighted measure of four prosodic features (f0, f0-slope, duration, and power) to measure the closeness between a given pair of phonemically equivalent database segments. Thresholds for the combination of these acoustic and prosodic measures have been determined by experience. The speech database is pruned by excluding all phone-sized segments that are within a given threshold of distance from another similar segment in terms of both left and right biphone contexts. By adjusting the pruning thresholds, we can determine the efficiency of the resulting speech database along a size-quality continuum. Smaller distance thresholds will produce a larger but more finely-graded corpus, while relaxed thresholds will further reduce the size of the corpus, though possibly at the expense of resulting synthesis quality

5. DATABASE DESIGN

There is an inherent problem with raw-waveform concatenative synthesis if it is to be used for the production of speech from unlimited input, since the coverage requirements of the source-unit database cannot easily be determined. However, if the task is finite, such as weather-forecasting, stock-announcements, traffic broadcasts, or online-trading, etc., then the required phonemic and prosodic combinations can be calculated in advance (or whenever new items are added) and the coverage of the database can be guaranteed. For tasks such as news-reading or proof-reading, however, the content cannot be predefined and coverage must be determined statistically from the analysis of representative text corpora [15, 16, 17].

Because most of the sentences used for recording a database of source-units have traditionally been gathered by greedy reduction from a very large written text corpus, they typically contain little interactive or conversational material suitable for the synthesis of dialogue speech. Most notably, they tend to be lacking in interrogative forms, so the speech which can best be synthesised from them is more suited to a formal reading style than to daily speech.

However, as more speech information is included in the synthesised voice, so the listener is likely to interpret more from that voice with respect to the speaker's intentions and beliefs. An interesting example (in Japanese) can be found in a synthesised reading of a weather forecast using a human voice that 'sounds sad'; the implications reach far beyond the content of the text (see [13]). While a neutral interpretation of the text is probably the most appropriate for tasks such as waether forecasting or news reading, where even in real-life situations the speaker is often not the originator of the text, this is not the case when the machine is speaking in place of a person and is required to represent not just the text of the utterance, but the mode of speaking as well.

This third category of speech synthesis use, as a communication-aid or talking device, for speech translation or for the orally handicapped, is becoming more important as the voice-quality of the synthesisers improves. For these devices, an additional 'voice-quality' or 'speaking-style' control is required in order to produce appropriate expressive speech for daily interactions.

Recent experience with Iida [18] who is using CHATR [12] to create a speech database for an ALS patient revealed many of the weaknesses of current synthesis approaches. The patient will soon lose the ability to speak and therefore wishes to preserve his present voice, which expresses his personality effectively, by the use of concatenative raw-waveform synthesis. The system is capable of using his voice to speak, but the texts that we provided him to read will allow only a minimum of coverage when considered against the needs of his daily life. In order to be expressive, they will have to include a range of emotions, a variety of question intonations, a full set of hesitations and fillers, and many other non-speech noises such as laughs and grunts that make human interactive speech meaningful.

Speech data collected from read texts will not be rich enough in variety to represent human conversational speech, and this presents an interesting challenge for database design.

6. CONCLUSION

This paper has described some tools and techniques for use in the creation of corpora for raw-waveform concatenative speech synthesis and presented a view of the trends at the turn of the century. It has stressed the need for balanced and representative source-unit databases for synthesis, and reflected on the paradigm shift wherein the size of the corpus greatly increases computing memory requirements while at the same time reducing the computing load, replacing rules with data.

We can foresee that cheaper memory will allow for even more data to be integrated into the synthesis process, and that the challenges of the coming years will be to define the coverage of that data such that it can be collected economically and yet represent the full range of human speech information. The linguistic coverage of current speeech synthesis will have to be matched by para-linguistic and extra-linguistic coverage so that the technology can find a place in the everyday lives of ordinary people.

With synthesis by concatenation of speech waveform segments, the technology has become extremely simple, reducing the synthesiser to a search-engine, but the design of the data and the features used to represent it in the index have become increasingly important. The task now is no longer to describe the basic elements of speech, such as vowel formant targets or prosodic contours, but to define the contexts by which these elements are controlled, in order to design a database with optimal coverage.

REFERENCES

[1] Campbell, W. N., "Talking machines for information acess", Tech Rept IEICE, 1999.

[2] Desirazu & Campbell, "An Extensible Scripting Interface for CHATR", Proc Acoust Soc Japan, Spring Meeting, pp309-310 1999.

[3] Snack http://www.speech.kth.se/snack/

[4] Allen, J., Hunnicutt, M. S. & Klatt, D.H. (1987), "From text to speech. The MITalk system", Cambridge University Press, Cambridge UK, 216 pages.

[5] Olive, J.P. (1980), "A scheme for concatenating units for speech synthesis", Proc. IEEE-ICASSP80, 568-571.

[6] Sagisaka, Y. (1988), "Speech synthesis by rule using an optimal selection of nonuniform synthesis units", Proc. IEEE-ICASSP88, 679-682.

[7] T. Hirokawa, "Speech synthesis using a waveform dictionary", pages 140-143, Proc Eurospeech, 1989.

[8] W.N.Campbell, "Labelling an English speech database for prosody control", 1-P-8, Proc ASJ, Spring, 1992.

[9] www.itl.atr.co.jp/chatr

[10] www.bell-labs.com/project/tts/sable.html

[11] W. N. Campbell & T. Marumoto, "Automatic labelling of voice-quality in speech databases for synthesis", these proceedings.

[12] W.N.Campbell, A.W.Black, "CHATR: Multilingual Speech Synthesis", IEICE Technical Report, SP96-7 1996.

[13] www.itl.atr.co.jp/chatr/j_tour/fkt_tenki.html

[14] Chen, J. D., & Campbell, W. N., "Speech Synthesis Evaluation by Objective Distance Measures", in SP-99-xxx, Tech Rept of the IEICE, May 1999.

[15] Campbell, N., & Saenko, E., "Factors to Consider in the Design of an Optimal Speech Corpus for Concatenative Speech Synthesis". Proc ASJ, Mar 1999.

[16] S. Deligne, F. Yvon, and F. Bimbot. Introducing statistical dependencies and structural constraints in variable-length sequence models. In *Grammatical Inference : Learning Syntax from Sentences*, Lecture Notes in Artificial Intelligence 1147, pages 156–167. Springer, 1996.

[17] Kawai, H., .Yamamoto, S., Higuchi, N., & Shimizu, T., "A design method of speech corpus for text-to-speech synthesis taking account of prosody" in these proceedings

[18] Iida, A., Campbell, N., Iga, S., Higuchi, I, & Yasumura, M., "Acoustic nature and perceptual testing of a corpus of emotional speech", Proc ICSLP-98.