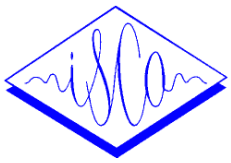

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

SPEAKEZ: A FIRST EXPERIMENT IN CONCATENATION SYNTHESIS FROM A LARGE CORPUS

Alexander G. Hauptmann

Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

This paper reports on a preliminary implementation of the SpeakEZ speech synthesis system. The system is built to explore the use of naturally occuring pronunciation, coarticulation and prosody for speech synthesis. SpeakEZ uses concatenative synthesis, choosing target phonemes from a database corpus of 115,000 prerecorded phonemes. The database of phonemes is segmented and labeled using the Sphinx speech recognition system. During synthesis, target phonemes are selected based on heuristics relating to phoneme context and syllable, word, and utterance position. The phonemes are concatenated in the time domain, using pitch synchronous overlap-add (PSOLA) smoothing between adjacent phonemes. Results from a preliminary evaluation of the system show that the system can at times provide excellent synthetic speech, but still has several shortcomings.[1]

KEYWORDS: Speech synthesis, time domain concatenation, corpus-based, pitch synchronous overlap-add (PSOLA).

1. INTRODUCTION AND RELATED RESEARCH

Speech synthesis allows the generation of spoken utterances from text. Synthesis is desirable when a large number of utterances must be available or when message content is unpredictable, requirements that make pre-recording of speech impractical. While major improvements have made commercial speech synthesizers highly intelligible, they still lack the naturalness found in normal human speech. One approach to getting more natural sounding synthetic speech is to use actual recorded human voices. However, since we can only record a limited number of utterances, we must concatenate pieces from a limited number of recordings to form a complete utterance. The standard approach to this problem so far has been to create an inventory of a few dozen to over a thousand carefully recorded and processed phonemes, diphones or demi-syllables [9, 3, 14] and to use those as the building blocks of phrases. The SpeakEZ system extends this approach by using more than 100,000 pre-recorded phonemes spoken by one speaker in many different contexts.

The problem then becomes one of composing a natural sounding utterance by selecting appropriate pieces of recordings, splicing them together and smoothing discontinuities. In addition, care must be taken to produce proper prosodics for the complete phrase.

One major difference in concatenative synthesis approaches has been the decision of whether to modify the selected recordings in the time domain or the frequency domain. Most approaches(e.g. [17]) have chosen to use an LPC-coded representation to allow easier modification of the prosodic parameters (pitch, duration, energy) to produce the phrase-level prosodic targets. An alternative is the PSOLA technique, which has recently been developed by researchers at CNET [10, 5], who use time domain modifications within a concatenative synthesis approach. At the heart of this approach is an analysis window, which is centered over each pitch peak in the source signal. The sequence of windowed analysis signals is then rescaled to provide the time scale and target pitch spacing desired. Further details of the PSOLA technique can be found in [11, 2]. We have decided to use the time domain PSOLA approach in the SpeakEZ system, which allows us to use the recorded samples without any further processing or distortion.

The SpeakEZ approach presented here also draws on work done by Nakajima and Hamada [12] as well as Sagisaka [16]. Each of these approaches pursues the basic idea of a large available corpus of recorded speech, from which the selection of target speech units is made.

The SpeakEZ system extends these approaches by using a much larger inventory of phonemes than any other system. It also exploits naturally occurring coarticulation phenomena and prosodic patterns present in the pre-recorded inventory. To manage this large phoneme inventory, a speech recognizer is used to automatically segment and label a database of recordings without human intervention.

2. THE SPEAKEZ SYSTEM

The SpeakEZ system is a preliminary implementation of a concatenative speech synthesizer. The SpeakEZ system was built to explore the possibility of using large amounts of recorded human speech for speech synthesis. The premise of the system is that if enough speech is recorded and catalogued in a database, then the synthesis consists merely of selecting the appropriate elements of the recorded speech and pasting them together. If we can collect enough instances of actual speech, then it should be possible to recreate the proper pronunciations and prosodics for any desired synthetic utterance by indexing into the recordings and splicing the appropriate elements together.

To see how well this strategy works, a listening experiment was conducted to estimate the intelligibility of the resulting utterances, and to allow comparison to other synthesis systems, for which published results on the same listening task are available.

SpeakEZ uses concatenation of phonemes from a large database

[1]This research was supported in part by the National Science Foundation Grant Number MDR-9154059; in part by the Advanced Research Projects Agency, ARPA Order 7239, monitored by the Space and Naval Warfare Systems Command under contract N00039-91-C-0158. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the United States Government.

of recorded speech in the time domain to obtain the target speech. Large databases of recorded speech can capture naturally occurring pronunciation and prosody, without the need for articulatory and prosodic modeling. At present, the system can produce a complete synthesized utterance from processed text, but is still missing the functionality of several important components. Most noticeably, there are no modules that determine prosodic targets or modify phrase level prosodics. As a result, the duration and intonational contour of elements in the utterance is frequently incorrect.

2.1 The database

The SpeakEZ system uses a database of over 115,000 recorded phonemes in a phonetically balanced corpus of over 3200 sentences. The database corpus includes 2333 TIMIT sentences [6], 720 "Harvard Corpus" sentences [4] and 200 sentences from the "Haskins Corpus" [13]. The corpus sentences were recorded from a single speaker over a period of several weeks. The creation of the database is completely automatic and no tuning or manual processing is required. This makes it feasible to exploit large amounts of stored data in speech synthesis, which previously had been prohibitive because of the human effort involved.

2.2 Segmentation and Labeling

The corpus utterances are phonetically labeled and categorized by the phonetic environment in which the phonemes were originally spoken. This is achieved by applying CMU's Sphinx automatic speech recognition system [7] in "forced segmentation" mode to each recorded sentence and its orthographic transcription. Since the phoneme segments are always used in a context of similar phonemes (see below), the segmentation procedure can tolerate small alignment errors introduced by the automatically segmenting recognizer, as long as they are consistent within the same phonetic context.

The database stores the following information about each phoneme:

- phoneme **class**,
- **source** utterance,
- **stress markers** associated with this phoneme (major, minor, unstressed),
- the phoneme **boundary** information (begin and end times according to the Sphinx recognition alignment),
- the identity of the **left and right context phonemes**,
- the relative position of this phoneme within the current **syllable**,
- the relative position of the phoneme within the current **word**,
- the relative position of the phoneme within the current **utterance**,
- and **pitch peak locations** as determined by a neural-net based pitch tracker [18].

2.3 Selecting Target Phonemes

During synthesis, input text is converted to phoneme strings using a standard text-to-phoneme algorithm originally developed for the MITALK system [1]. The database is searched for phonemes that appear in the same context as the target phoneme string. A penalty for the context match is computed as the difference between the immediately adjacent phonemes surrounding the target phoneme and the corresponding phonemes adjacent to the database phoneme candidate. This penalty is based on the difference between 16 phonetic features of two phonemes. The context match is also influenced by the distance of the phoneme to its left and right syllable boundary, left and right word boundary, and to the left and right utterance boundary. There are many ways in which this heuristic could be modified and improved. However, for the evaluation experiments described below, the heuristic was applied without changes.

Specifically, the current heuristic operates the following way: For each phoneme i in the database that is of the same type as the target phoneme j, we compute a score which is the sum of the following components:

- StressPenalty * (difference in stress between DB phoneme(i) and the target phoneme(j))
- LeftContextPenalty * (acoustic feature similarity between DB phoneme(i-1) and the target phoneme(j-1)).
- RightContextPenalty * (acoustic feature similarity between DB phoneme(i+1) and the target phoneme(j+1)).
- LeftContextStressPenalty * (difference in stress between DB phoneme(i-1) and the target phoneme(j-1))
- RightContextStressPenalty * (difference in stress between DB phoneme(i+1) and the target phoneme(j+1))
- LeftSyllablePositionPenalty * (difference in relative position between DB phoneme(i) and target phoneme(j) from the beginning of the syllable)
- RightSyllablePositionPenalty * (difference in relative position between DB phoneme(i) and target phoneme(j) from the end of the syllable)
- LeftWordPositionPenalty * (difference in relative position between DB phoneme(i) and target phoneme(j) from the beginning of the word)
- RightWordPositionPenalty * (difference in relative position between DB phoneme(i) and target phoneme(j) from the end of the word)
- LeftUtterancePositionPenalty * (difference in relative position between DB phoneme(i) and target phoneme(j) from the beginning of the utterance)
- RightUtterancePositionPenalty * (difference in relative position between DB phoneme(i) and target phoneme(j) from the end of the utterance)

Based on informal experiments with these features, we established that the rank order of penalty weighting should be StressPenalty >> Right and LeftContextPenalty >> Right and LeftWordPenalty = Right and LeftSyllablePenalty >> Right and LeftUtterancePenalty. The phoneme with the lowest penalty score is selected as the target phoneme for concatenation.

2.4 Synthesis

For each phoneme in the target utterance to be synthesized, the best matching phoneme (with the minimal penalty score) from the database is chosen and spliced into the output speech. Energy levels are smoothed between phoneme boundaries. The pitch periods are also adjusted in the time domain using the *Pitch Synchronous OverLap Add* (PSOLA) technique [5, 2]. Since we have no module to predict target values for the pitch and energy of the target phoneme, the PSOLA algorithm is only used to smooth pitch transitions between phonemes within a window of three (3) pitch periods. The resulting speech has many

coarticulation and prosodic cues preserved from the natural examples in the database.

This implementation of the SpeakEZ system reflects the conceptual approach stated earlier. A large database of phonemes is exploited for naturally occurring phenomena related to coarticulation and prosody. Rather than modify existing prosody according to ideal target values, the system uses the exact duration, intonation and articulation of the database phoneme without modifications. The selection heuristic attempts to find a database phoneme which best matches the target phoneme according to criteria of context as well as position relative to the nearest word and syllable boundaries.

3. EVALUATION

Three separate experiments were used to estimate the intelligibility of the SpeakEZ synthesizer.

The first test, known as the *Modified Rhyme Test* (MRT) [8] determined the segmental intelligibility of single words. In the MRT, the listener is presented with the sound for a single monosyllabic word and is forced to choose between six response alternatives for the word. E.g., when the word *fill* is played, the listener is given the choice of selecting either *fill, fig, fin, fizz, fib* or *fit* as the appropriate response selection. While this particular set of alternatives only tests the intelligibility of the final consonants on the word, a similar response set for the initial consonant identification is also possible, asking the listener to choose between *fill, will, hill, kill, till* and *bill.*

The second intelligibility test corpus is known in the literature as the *Haskins semantically anomalous sentences* [8]. This corpus consists of a set of sentences which were specifically designed to evaluate synthetic speech utterances. Sentences in the Haskins corpus all have the form *"The <adj> <noun> <verb> the <noun>"* where any adjective can appear as *<adj>*, any noun could be *<noun>* and any verb can take the role of *<verb>*. Examples of these sentences are: *"The wrong shot led the farm."* and *"The salt dog caused the shoe."*

The final test corpus was taken from the *Harvard sentence corpus.* The Harvard sentences are both semantically and syntactically meaningful sentences that comprise a phonetically balanced corpus [4]. Examples of such sentences are: *"The boss ran the show with a watchful eye"* and *"Clothes and lodging are free to new men."* Listening recognition of this final corpus tends to be higher since the meaning of sentences provides strong clues to insufficient acoustic evidence.

Even though some of these sentences had originally been used in the construction database of the SpeakEZ system, their respective database entries were removed to avoid testing intelligibility on a sentence that had been completely recorded as part of the database construction.

Subjects for each evaluation were recruited among the staff and students of the Carnegie Mellon University. All subjects had some background in speech recognition and speech synthesis research. Three listening tests were conducted, one for the MRT listening task, one for Haskins listening task and one for the Harvard listening task. Each test was conducted with five subjects, each subject listening to a sample of 50 randomly selected utterances from the respective task corpus. All subjects were native English speakers and none reported any hearing deficiencies. Each subject could listen to a phrase, decide on an appropriate response and then type in a transcription to the computer. Subjects were also allowed to replay a recording if external events interrupted the playback of an utterance. Subjects were allowed to proceed at their own pace, moving on to the next sentence when ready. This was designed to mirror the experimental design described in [8, 15] to allow comparisons with those results.

System	MRT	Haskins	Harvard
SpeakEZ	88.7	92.1	95.5
Natural Speech	99.4	97.7	99.2
DECTalk (Paul)	96.7	86.8	95.3
DECTalk (Betty)	94.4	75.1	90.5
Mitalk-79	93.1	78.7	93.3
Prototype PROSE-2000	87.6	64.5	83.7

Table 1. Percent Correct Word Recognition on 3 Listening Tasks. The MRT task tests single word intelligibility, the Haskins task requires comprehension of syntactically correct but semantically anomalous sentences, and the Harvard task tests intelligibility of meaningful sentences. The data for all systems except SpeakEZ are from Pisoni et al., 1985 [15]. These data can only be used for approximate comparisons, due to different subject pools and experimental procedures.

3.1 Results

All error rates were computed by adding the total number of word deletions, word substitutions and word insertions, and dividing by the total number of words in the target sentences. The average word error rate in the MRT listening experiments was 11.3 percent. For the Haskins corpus (semantically anomalous, syntactically correct) listening experiments, the average word error rate was 7.9 percent. In the Harvard corpus (syntactically correct, semantically correct) the average word error rate was 4.5 percent.

To put these results in perspective, Table 1 lists some numbers reported in [8, 15] that should be comparable since they were obtained from the same listening tasks.

But the results should be regarded with caution. They are derived from a small sample size of speech researchers, not from the general public as in [15]. The Haskins listening test could also be considered biased for the following reason: Haskins sentences were included in the database of recordings used to build the speakEZ system. Even though the specific matching sentences were excluded from the database during the intelligibility evaluation, other sentences with the some of the same words and identical sentence structure were used to construct the answers. Thus the performance on the Haskins sentences is closer to an upper bound performance than a true listening test evaluation of the capabilities of the SpeakEZ system.

4. CONCLUSIONS

In this paper we have described an initial implementation of a speech synthesis approach. Our evaluations have been encouraging, but they still fall short of the ultimate goal of speech that is indistinguishable from human speech. The SpeakEZ system suffers from several major deficiencies and future work is outlined below to remedy some of these shortcomings.

The system requires fairly large data storage facilities. The uncompressed data files for 3253 sentences in the corpus use about 360 MB of data sampled in 16 bit samples at 16Khz. While this is a relatively large amount of data, current technology lets us imagine even larger data sets using new storage media, such as CD-ROM. Thus we feel further exploration of this approach

is justified, despite the apparent inefficiency due to the massive data requirements.

The premise of the system was to exploit the use of naturally occurring coarticulation and prosodic patterns in the recorded database. This proved to be successful, since the system produced intelligible speech without resorting to any rules for prosody other than the target phoneme selection heuristic. In fact, the system at times produced output nearly indistinguishable from natural speech, but occasionally these heuristics broke down, resulting in barely intelligible portions of some utterances.

The most glaring shortcoming of the system is the lack of proper prosodic target information in the synthesizer. While the heuristics for selecting a reasonable phoneme from the database are frequently quite adequate, occasionally they are insufficient. In this case, the resulting speech sounds very awkward, and unnatural. Both duration modeling and phrase-level intonation contours would help the system in overcoming these shortcomings.

There are also cases where the segmentation produced by the Sphinx recognizer is not adequate for the SpeakEZ system. We hope to use an improved version to obtain more consistent and accurate segmentation in the future.

The results should not convey the notion that the SpeakEZ system is better than DecTalk. It is not. The results on the MRT test are probably more relevant to indicate its performance relative to DecTalk and other synthesizers. However, the results are very good considering that SpeakEZ is a system with no special knowledge of duration, intonation, syntax and semantics. It only has the information described above in the database, and it exploits that with surprising success. We believe the system is a promising approach to high quality synthesis and these initial pilot results are a clear indication of the potential.

Future work on the SpeakEZ system will focus on better modeling of phoneme and word durations during the creation of the target phoneme string. Phrase level prosodic features will also be added as labelled features into the database information. These features will emphasize prosodic information, such as phrase boundaries, function words, and pitch contours.

We have found that speech synthesis by concatenation from a large corpus of recordings is extremely promising. However, our results indicate that we still have many improvements ahead of us before we can rival natural speech in intelligibility.

REFERENCES

[1] ALLEN, J., CARLSON, R., GRANSTROEM, B., HUNNICUTT, S., KLATT, D., AND PISONI, D. *Conversion of Unrestricted English Text to Speech.* Cambridge, MA, 1979.

[2] CHARPENTIER, F., AND MOULINES, E. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Proceedings Eurospeech 89* (Paris, 1989), pp. 13 – 19.

[3] CHARPENTIER, F., AND STELLA, M. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In *ICASSP 86* (Tokio, 1986), IEEE, pp. 2015 – 2018.

[4] EGAN, J. Articulation testing methods. *Laryngoscope 58* (1948), 955 – 991.

[5] HAMON, C., MOULINES, E., AND CHARPENTIER, F. A diphone synthesis system based on time-domain prosodic modifications of speech. In *ICASSP 89* (1989), IEEE, pp. 238 – 241.

[6] LAMEL, L., KASSEL, R., AND SENEFF, S. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings Speech Recognition Workshop* (1986), DARPA, pp. 100 – 109.

[7] LEE, K.-F. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The Sphinx System.* PhD thesis, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, 1988. Tech Report CMU-CS-88-148.

[8] LOGAN, J., GREENE, B., AND PISONI, D. Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America 86*, 2 (August 1989), 566 – 581.

[9] MACCHI, M., AND SPIEGEL, M. Using a demi-syllable inventory to synthesize names. In *Speech Tech '90, Official Proceedings, Voice Input/Output Applications, conference and exhibition* (New York, NY, April 1990), vol. 2, Media Dimensions, Inc., pp. 208 – 212.

[10] MOULINES, E., EMERARD, F., LARREUR, D., LE SAINT MILON, J., LE FAUCHEUR, L., MARTY, F., CHARPENTIER, F., AND SORIN, C. A real time french text to speech system generating high quality synthetic speech. In *ICASSP-90, Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (Albuquerque, NM, April 1990), vol. 1, IEEE.

[11] MOULINES, E., HAMON, C., AND CHARPENTIER, F. High-quality prosodic modifications of speech using time-domain overlap-add synthesis. In *Douzieme Colloque GRETSI* (Juan-les-Pins, 1989), pp. 541 – 544.

[12] NAKAJIMA, S., AND HAMADA, H. Automatic generation of synthesis units based on context oriented clustering. In *ICASSP-88, Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (New York, April 1988), vol. 1, IEEE, pp. 659 – 662.

[13] NYE, AND GAITHENBY. The intelligibility of synthetic monosyllabic words in short syntactically normal sentences. Tech. Rep. SR 37/38, Haskins Laboratory, 1974.

[14] OLIVE, J. Speech synthesis by rule. In *Speech Communication*, G. Fant, Ed., vol. 2. J. Wiley, New York, NY, 1974. Proceedings of the speech communication seminar in Stockholm, Aug 1 - 3, 1974.

[15] PISONI, D., NUSBAUM, H., AND GREENE, B. Perception of synthetic speech generated by rule. *Proceedings of the IEEE 73*, 11 (1985), 1665 – 1676.

[16] SAGISAKA, Y. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *ICASSP-88, Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (New York, April 1988), vol. 1, IEEE, pp. 679 – 682.

[17] SPIEGEL, M. Orator system technical briefs no.1. Tech. rep., Bell Communications Research Labs, March 1991.

[18] ZHOU, L. *A Speaker-Independent Neural Network Pitch Tracker with Telephone Bandwidth Speech for Computer Speech Recognition.* PhD thesis, CSE Department, Oregon Graduate Institute, February 1991. Technical Report CS/E 91-TH-002.