# Corpus-Based Unit Selection TTS for Hungarian

Márk Fék, Péter Pesti, Géza Németh, Csaba Zainkó, and Gábor Olaszy

Laboratory of Speech Technology
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Hungary
{fek, nemeth, zainko, olaszy}@tmit.bme.hu, pesti@alpha.tmit.bme.hu

**Abstract.** This paper gives an overview of the design and development of an experimental restricted domain corpus-based unit selection text-to-speech (TTS) system for Hungarian. The experimental system generates weather forecasts in Hungarian. 5260 sentences were recorded creating a speech corpus containing 11 hours of continuous speech. A Hungarian speech recognizer was applied to label speech sound boundaries. Word boundaries were also marked automatically. The unit selection follows a top-down hierarchical scheme using words and speech sounds as units. A simple prosody model is used, based on the relative position of words within a prosodic phrase. The quality of the system was compared to two earlier Hungarian TTS systems. A subjective listening test was performed by 221 listeners. The experimental system scored 3.92 on a five-point mean opinion score (MOS) scale. The earlier unit concatenation TTS system scored 2.63, the formant synthesizer scored 1.24, and natural speech scored 4.86.

## 1 Introduction

Corpus-based unit selection TTS synthesis creates the output speech by selecting and concatenating units (e.g. speech sounds or words) from a large (several hours long) speech database [1]. Compared to TTS systems using diphone and triphone concatenation, the number of real concatenation points becomes much smaller. Moreover, the database of traditional diphone and triphone concatenation TTS systems is recorded with monotonous prosody whereas the units from a large speech corpus retain their natural and varied prosody. Thus, it becomes possible to concatenate larger chunks of natural speech, providing superior quality over diphone and triphone concatenation.

Corpus-based unit selection TTS systems have already been developed for several major languages. The state-of-the-art Hungarian TTS systems use diphone and triphone concatenation, see for example [2]. These systems allow unrestricted domain speech synthesis, but the speech quality is limited by the non-unit-selection based waveform concatenation technology. In this paper, we describe our ongoing work in developing a corpus-based unit selection TTS for Hungarian. Our first goal was to develop a restricted domain TTS capable of reading weather forecasts. Based on the experience gained, we plan to extend the system to read unrestricted texts.

Section 2 describes the text collection and the design of the text corpus. Section 3 details the recording and labeling of the speech database. Section 4 describes the mechanism of the unit selection. Finally, Section 5 describes the result of a subjective evaluation test comparing the quality of the system to that of earlier Hungarian TTS systems.

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 367–373, 2006.

## 2 Text Collection and Corpus Design

We collected texts of daily weather forecasts in Hungarian from 20 different web sites for over a year. After spell checking and resolving abbreviations, the resulting text database contained approximately 56, 000 sentences composed of about 493, 000 words (5, 200 distinct word forms) and 43, 000 numbers. Almost all of the sentences were statements; there were only a few questions and exclamations. On the average, there were 10 words in a sentence (including numbers). The average word length was slightly over 6 letters because of the frequent presence of longer than average weather related words. Statistical analysis has shown that as little as the 500 most frequent words ensured 92% coverage of the complete database, while the 2, 300 most frequent words gave 99% coverage. The next paragraph describes a more detailed analysis that takes the position of the words within a prosodic phrase into consideration. We have obtained similar results on data collected over only a half year period, thus we assume that these results are mainly due to the restricted weather forecast domain. As Hungarian is an agglutinative language, a corpus from an unrestricted domain requires approximately 70, 000 word forms to reach 90% coverage [3]. The favorable word coverage of the restricted domain allowed us to choose words as the basic units of the speech database.

The next step was to select a part of the speech corpus for recording. We used an iterative greedy algorithm to select the sentences to be read. In each iteration, the algorithm added the sentence that increased word coverage the most. A complete coverage was achieved with 2, 100 sentences. We extended the algorithm to include some prosodic information derived from position data. Sentences were broken into prosodic phrases, using punctuation marks within a sentence as prosodic phrase boundaries. We assigned two positional attributes to each word: the position of the word within its prosodic phrase, and the position of the prosodic phrase (containing the word) within the sentence. Both positional attributes may have three values: first, middle, or last. Our underlying motivation was that words at the beginning and at the end of a prosodic phrase tend to have a different intonation and rhythm than words in the middle of the prosodic phrase. Similarly, the first and the last prosodic phrases in a sentence tend to have a different intonation than the prosodic phrases in the middle of a sentence. We obtained 5, 200 sentences containing 82, 000 words by running the extended algorithm. The sentences contain 15, 500 words with distinct positional attributes and content.

A speech synthesizer using words as units can only synthesize sentences whose words are included in the speech database. In a real application it may occur that words to be synthesized are not included in the database. In order to synthesize these missing words, we have chosen speech sounds to be the universal smallest units of the speech database. As we use speech sounds only as an occasional escape mechanism to synthesize words not included in the speech database, we did not optimize the database for triphone coverage.

## 3 Speech Database Recording and Labeling

The selected sentences were read by a professional voice actress in a sound studio. We have also recorded 60 short sentences covering all the possible temperature values. The speech material was recorded at 44, $1kHz$ using 16 bits per sample. The recording sessions spanned over four weeks with $2 - 3$ days of recording per week and $4 - 5$ hours of recording per

day. The recorded speech material was separated into sentences. The 5260 sentences resulted in a database containing 11 hours of continuous speech. We have extracted the fundamental frequency of the waveforms by using the autocorrelation based pitch detection implemented in the Praat software [4]. Pitch marks were placed on negative zero crossings to mark the start of pitch periods in case of voiced speech and at every $5ms$ in case of unvoiced speech. When concatenating two speech segments, the concatenation points are restricted to be on pitch marks, assuring the phase continuity of the waveform. The fundamental frequency itself is not stored but recalculated from the pitch periods when needed.

The word and speech sound unit boundaries are marked automatically. To mark the sound boundaries in the speech waveform, a Hungarian speech recognizer was used in forced alignment mode [5]. We performed a manual text normalization by expanding numbers, signs, and abbreviations in the textual form of the sentences before and during the recording. The speech recognizer performs an automatic phonetic transcription. The phonetic transcription inserts optional silence markers between words, and takes into account the possible co-articulatory effects on word boundaries. Thus, it provides a graph of alternative pronunciations as input to the speech recognizer. The speech recognizer selects the alternative that best matches the recorded speech, and also returns the corresponding phonetic transcription. The hidden Markov model based speech recognizer was trained with a context dependent triphone model on a Hungarian telephone speech database [6]. The preprocessing carries out a Mel cepstral (MFCC) analysis using a fixed frame size of $20ms$ and a frame shift of $10ms$. The detected sound boundaries are aligned to the closest pitch mark.

We performed a statistical analysis on sound durations using the detected sound boundaries and manually checked sounds with extreme durations. We identified and corrected several sentences where the waveform and the textual content of the sentence did not match, due to mistakes in the manual processing of the database. Apart from that, we have observed some problems concerning the incorrect detection of sound boundaries for unvoiced fricatives and affricates. The problem is likely caused by the use of telephone speech to train the recognizer, because telephone speech does not represent frequencies above $3400Hz$ where unvoiced fricatives and affricates have considerable energy. We plan to correct the problem by retraining the recognizer on the recorded 11 hour long speech corpus.

The word boundaries were marked automatically on the phonetic transcription returned by the recognizer. Separate markers were used for identifying the beginning and the end of each word. Each marker was assigned to a previously detected sound boundary. In some cases, the last sound of a word is the same as the first sound of the following word. If there is a co-articulation effect across word boundaries, only one sound will be pronounced instead of two. In this case, we place the word boundary end/start markers after/before the fusioned sound so as to include the sound in both words. When selecting the waveform corresponding to words starting/ending with such speech sounds, only 70% of the fusioned sound is kept, and their first/last 30% is dropped.

## 4 Unit Selection

The unit selection algorithm divides the input into sentences and processes every sentence separately. The algorithm follows a two-phase hierarchical scheme [7] using words and speech sounds as units. In the first phase of the algorithm, only words are selected. If a

word is missing from the speech database, the algorithm composes it from speech sounds in the second phase. The advantage of the hierarchical scheme is that it makes the searching process faster. We plan to add an intermediate, syllable level to the system, which may work well in case of unrestricted domain synthesis.

The unit selection algorithm identifies the words based on their textual content. A phonetic transcription is also generated and used for identifying the left and right phonetic context of the words. The speech sounds are identified by the phonemes in the phonetic transcription. A list of candidate units with the same textual (or phonetic) content is created for every word (or speech sound) in the sentence (or word) to be synthesized.

The unit selection algorithm uses two cost functions. The target cost captures how well a unit in the speech corpus matches a word (or speech sound) in the input. The concatenation cost captures how natural (or smooth) the transition is between two concatenated units sounds. The number of candidates for a given unit is limited to reduce the search time. If there are more candidates than the limit, only the ones with the lowest target cost are kept. The Viterbi algorithm is used to select the optimum path among the candidates giving the smallest aggregated target and concatenation cost.

In our implementation, the target cost is composed of the following subcosts:

1. The degree of match between the left and right phonetic contexts of the input unit and the candidate. This part of the target cost is zero, if the phonetic contexts are fully matched. We have defined seven phoneme classes for consonants, based on their place of articulation (bilabial, labiodental, dental, alveolar, velar, glottal, nasal) [8]. Consonants within the same class tend to have similar co-articulation effects on neighboring sounds. The target cost is smaller for phonemes in the same class, and becomes bigger if the preceding or following phonemes are from different classes. The target costs between the different phoneme classes are defined in a cost matrix. The weights in the matrix were set in an ad-hoc way. Further optimization may improve the quality of the system.
2. The degree of match between the position of the input word and the position of the candidate within their respective prosodic phrases. The positions can take three values: first, middle, or last. This subcost is only defined for words.
3. The degree of match between the relative positions of the prosodic phrases (containing the input word or the candidate) within their corresponding sentences. This subcost is only defined for words.

The concatenation cost is calculated as follows:

1. Units that were consecutive in the speech database have a concatenation cost of 0, because we cannot have a better concatenation than in natural speech. This motivates the algorithm to choose continuous speech segments from the database.
2. Candidates from the same database sentence have lower concatenation cost than candidates from different sentences. This gives a preference to concatenate units with similar voice quality.
3. Continuity of fundamental frequency ($F_0$), calculated as the weighted difference between the ending $F_0$ of the first and the starting $F_0$ of the second unit.

The various weights of the two cost functions were tuned manually during informal listenings, on test sentences not included in the corpus.

**Table 1.** Mean opinion scores per sentence. The confidence ($\alpha = 0.05$) took values between 0.04 and 0.09.

| sentence number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | variance |
|---|---|---|---|---|---|---|---|---|---|---|---|
| natural | **4.83** | **4.83** | **4.91** | **4.88** | **4.79** | **4.90** | **4.86** | **4.92** | **4.84** | **4.84** | 0.04 |
| corpus-based | **4.75** | **4.28** | **4.16** | **3.56** | **3.59** | **3.85** | **4.29** | **3.63** | **3.75** | **3.30** | 0.44 |
| diphone-triphone | **2.52** | **2.88** | **2.79** | **2.56** | **2.66** | **2.60** | **2.44** | **2.76** | **2.65** | **2.46** | 0.15 |
| formant synthesis | **1.25** | **1.20** | **1.33** | **1.26** | **1.25** | **1.20** | **1.26** | **1.21** | **1.21** | **1.20** | 0.04 |

## 5 Subjective Evaluation

We have carried out a subjective listening test to compare the quality of our corpus-based unit selection TTS system to that of a state-of-the-art Hungarian concatenative TTS system [2]. We have also included a Hungarian formant synthesizer [9] in the test to measure the evolution of quality across different TTS generations.

We decided to limit the length of the test to 10 minutes to make sure that the listeners do not lose their interest in the test. Listeners were asked to evaluate the voice quality of the synthetic speech after every sentence heard. Intelligibility was not evaluated, because we do not expect it to be a real problem for weather forecasts. The listeners had to evaluate the quality of the synthesized speech on the following 5-point scale: excellent (5), good (4), average (3), poor (2), bad (1).

The content of the test sentences was matched to the weather forecast application. We chose 10 random sentences from a weather report. The weather report originated from one of the web sites included in the database collection. Thus, the style of the sentences was close to the speech corpus, but the chosen sentences were not included in the corpus. A listener had to evaluate 40 sentences (10 natural, 10 generated by the formant synthesizer, 10 generated by the diphone-triphone synthesizer, and 10 generated by the corpus-based synthesizer) in a pseudo-random order.

The test was carried out via the Internet using a web-interface. This allowed the participation of a large number of test subjects. The average age of the 248 listeners was 22.9 years. Most of them were students. The results from 185 males and 36 females were evaluated, while 27 listeners were excluded because we judged their results as inconsistent. At the beginning of the test, the testers had to listen to an additional $11th$ weather report sentence in four versions. This allowed the listeners to familiarize themselves with the different speech qualities. Each sentence was played only once to reduce the length of the test. According to the listener responses to our questionnaire, most of them carried out the test in a quiet room using average quality equipment.

We excluded testers from further evaluations who gave an 'average (3)' or worse score to natural speech samples at least twice. We supposed that these excluded testers were either guessing, or had difficulty with the playback. According to our preliminary tests, the playback function did not work continuously for large speech files in case of slow Internet connections. Therefore we have converted all speech samples to $22kHz$ and compressed them with a $32kbps$ variable bit rate MPEG1-LIII encoder. We did an informal evaluation with high quality headphones and found no quality difference between the encoded and the original speech samples.

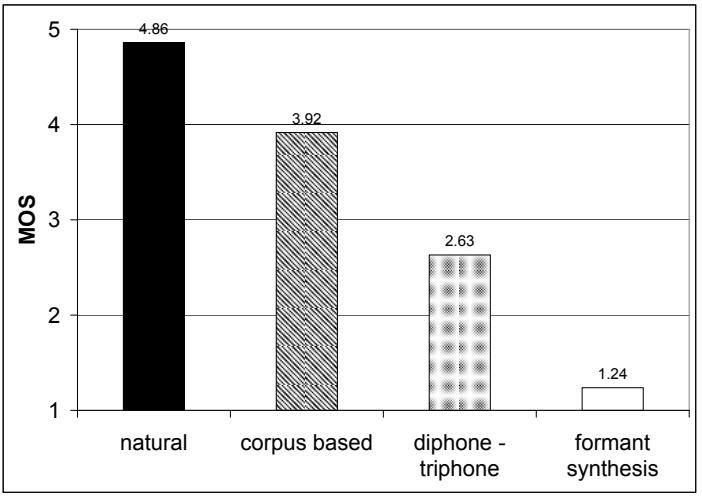


**Fig. 1.** Mean opinion scores obtained for the different TTS systems. The confidence ($\alpha = 0.05$) took values between 0.02 and 0.03.

**Table 2.** Relation of the MOS values to the number of real concatenation points in a sentence synthesized by the corpus-based system.

| sentence number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| MOS (corpus-based) | **4.75** | **4.28** | **4.16** | **3.56** | **3.59** | **3.85** | **4.29** | **3.63** | **3.75** | **3.30** |
| number of concatenation points | **3** | **4** | **4** | **6** | **9** | **10** | **12** | **14** | **15** | **24** |
| number of words | 10 | 11 | 8 | 7 | 10 | 12 | 9 | 15 | 25 | 22 |
| number of concatenated words | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

The resulting Mean Opinion Scores (MOS), summarized in Figure 1, show a major quality difference between the different synthesizers. The corpus-based synthesizer outperformed the diphone-triphone concatenation system by 1.3 points, which indicates that we may expect higher user acceptance and more widespread use of the corpus-based system.

We have explored the correlation between perceived quality and the number of real concatenation points in a synthesized sentence. We define the real concatenation point as a point separating two speech segments in the synthesized sentence that were not continuous in the speech database. Table 2 shows the sentences ordered by the number of concatenation points. The best MOS was achieved by the sentence containing the least, i.e. 3 concatenation points. The worst quality was achieved by the sentence containing the most, i.e. 24 concatenation points. The speech quality, however, does not depend consistently on the number of concatenation points. The $7th$ sentence, for instance, has the second best quality but contains more (12) concatenation points than most of the sentences. The correlation between the MOS scores and the number of concatenation points is $-0, 68$. Table 2 also shows, that there was only one sentence where it was necessary to use speech sounds as units.

## 6 Conclusion

In this paper, we have described our ongoing work on a corpus-based unit selection TTS system for Hungarian. We have built an experimental application for synthesizing restricted domain weather forecasts. The quality of the system was evaluated using a subjective listening test. 10 sentences from a weather forecast were synthesized by the corpus-based unit selection system. The sentences were also synthesized by two unrestricted domain TTS systems using non-unit-selection based diphone/triphone concatenation and formant synthesis. The new system outperformed the diphone/triphone concatenation by 1.3 MOS points, and the formant synthesis by 2.7 MOS points. The quality of the experimental TTS system showed a greater variance depending on the input sentence than the other two systems. Some correlation was found between the number of concatenation points in a sentence and its quality. We expect to further improve the quality by introducing fundamental frequency smoothing. Our future plan is to improve the prosody model and the unit selection algorithm to be able to extend the system to general unrestricted TTS synthesis.

## Acknowledgments

We would like to thank our colleagues, Mátyás Bartalis, Géza Kiss, and Tamás Bőhm for their varoius contributions. We also thank all the listeners for participating in the test.

This project was funded by the second Hungarian National R&D Program (NKFP), contract number 2/034/2004.

## References

1. Möbius, B.: Corpus-Based Speech Synthesis: Methods and Challenges. AIMS 6 (4), Univ. Stuttgart, pp. 87–116., 2000.
2. Olaszy, G., Németh G., Olaszi, P., Kiss, G., Gordos, G.: PROFIVOX - A Hungarian Professional TTS System for Telecommunications Applications. International Journal of Speech Technology, Volume 3, Numbers 3/4, December 2000, pp. 201–216.
3. Németh, G., Zainkó Cs.: Word Unit Based Multilingual Comparative Analysis of Text Corpora. Eurospeech 2001, pp. 2035–2038., 2001.
4. Boersma, P.: Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound. IFA Proceedings 17, pp. 97–110., 1993.
5. Mihajlik P., Révész T., Tatai P.: Phonetic Transcription in Automatic Speech Recognition. Acta Linguistica Hungarica, Vol. 49 (3–4), pp. 407–425, 2002.
6. Vicsi, K., Tóth, L. Kocsor, A., Gordos, G., Csirik, J.: MTBA - Magyar nyelvű telefonbeszéd adatbázis (Hungarian Telephone-Speech Database). Híradástechnika, vol. 2002/8., pp. 35–39, 2002.
7. Taylor, P., Black, A., W.,: Speech Synthesis by Phonological Structure Matching. Eurospeech 1999, vol. 2, pp. 623–626, 1999.
8. Olaszy, G.: Az artikuláció akusztikus vetülete – a hangsebészet elmélete és gyakorlata (The Articulation and the Spectral Content—the Theory and Practice of Sound Surgery). in: Hunyadi, L. (ed.): KIF-LAF (Journal of Experimental Phonetics and Laboratory Phonology), Debreceni Egyetem, pp. 241–254, 2003.
9. Olaszy, G., Gordos, G., Németh, G.: The MULTIVOX Multilingual Text-to-Speech Converter. in: G. Bailly, C. Benoit and T. Sawallis (eds.): Talking machines: Theories, Models and Applications, Elsevier, pp. 385–411, 1992.