Design issues of a corpus-based speech synthesizer

ANDRÁS NAGY, PÉTER PESTI, GÉZA NÉMETH, TAMÁS BŐHM

Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics

{nagy.a, pesti}@alpha.tmit.bme.hu

Reviewed

Key words: synthesized speech, speech quality, sampling, corpus volume

The corpus-based approach is a new technique which has never been used in Hungary. It offers a more flexible and better quality synthesis. This article outlines the basic principles of this technique then a more detailed description follows of the development of a Hungarian corpus-based, object-related system being under development at the Speech Research Laboratory of the Budapest University of Technology and Economics. In the second part of the article statistical studies with weather forecasts are introduced then some considerations regarding the selection of announcers are presented. Finally some other design issues of corpus-based systems are addressed.

1. Introduction

As a result of the convergence and integration of telecommunications, information and media technologies today, our world is headed towards the realization of an information based society. The process with the most profound effects in this transition – in addition to the advancement of networks, mobility and computers – is the transformation of human-computer interaction. Speech technologies, such as speech synthesis, play a fundamental role in this change.

Recent years have brought the new concept of corpus based speech synthesis [1]. The core idea of the concept is the generally accepted principle that the quality of a waveform concatenation based speech synthesizer will be determined largely by the number of concatenation points. As the length of the elements used in the synthesized speech increases, the number of concatenation points decreases, resulting in higher perceived quality.

In the ideal case, all possible texts, or at least all possible sentences would be stored as a single waveform element in the system's database. Since this is infeasible in a practical implementation, shorter units are introduced into the database, while aspiring to concatenate the output speech using few elements with high probability. The flexibility of the system makes it desirable to use variable length elements instead of a certain fixed length [2].

A number of speech synthesizers have been created for some languages spoken by many people following the ideas outlined above [2], but a Hungarian implementation was not yet available. The objective of our work is therefore to create such a modern speech synthesis system, building on results and experiences of previous Hungarian solutions (Profivox [3,4] and number synthesizer [5]). Since the creation of the required complex software is a task for several years, we are developing a closed domain synthesizer first for weather reports. This system allows simplified design, while serving as the foundation for a wide – or possibly unrestricted – domain synthesizer.

This article reports on the first phase of the research and development process. We give an overview of the primary challenges of corpus based synthesis, propose solutions to some of the problems, and present our initial experiments, which we use to evaluate the potential of the concept. The article is concluded by providing a summary of our work so far, and outlining the research and development tasks lying ahead.

2. Design challenges in corpus based systems

In this section, we give a brief insight into the challenges of designing corpus based systems, describe our experiments and possible solutions to some of the problems.

2.1. Speaker selection

Matching of waveform elements, originally cut from different parts of the speech corpus, is of paramount importance to the quality of the synthesized speech. This is ensured by an item selection algorithm, but the speaker's ability to produce speech with constant prosody properties predetermines the selection possibilities.

It is a basic requirement that the fundamental frequency (pitch) of the speaker should not fluctuate in a wide range, for example. Although the pitch can later be modified by signal processing methods, this intervention has an adverse effect on the quality of the synthesis. These considerations have led us to define some requirements to be met by a speaker.

These requirements were the following: clear articulation, pleasant tone, consistency (ability of the speaker to produce the same phonemes in a similar fas-

hion within one session, and also between sessions), and availability (sufficient amount of audio recordings accessible from the speaker). The eventual selection of a speaker was made in several steps based on the requirements.

We downloaded two full days of the broadcast archives for the stations available on the homepage of the Hungarian Radio (Kossuth, Bartók and Petőfi stations). The audio files were accessible in an hourly breakdown in RealAudio format, but the quality of these files didn't permit detailed acoustic analyses.

Characteristics of the speakers were collected by listening to the audio files multiple times. Comparison of these features with each other and with the initial requirements has resulted in our list of speakers deemed most appropriate for inclusion in the speech corpus.

We requested high quality audio recordings for the selected speakers from the internal archives of the Hungarian Radio. These files allowed more detailed investigations of acoustic properties, with the pitch and intensity as the most important aspects. We studied the values of these features on the time scale, the averages of the values, and the deviation from the average. The analysis was concluded by proposing the speaker with the most advantageous characteristics.

2.2. Issues of element selection

The key idea of corpus based synthesis is the availability of multiple versions of elements for concatenation during synthesis, making the selection of the best element possible according to a given metric. While in a diphone synthesizer the only consideration is the match of the phonetic labels of concatenated diphones, in the corpus based solution multiple aspects can be balanced with the use of a compound cost function.

The metric describing the correspondence between a selected element and the portion of speech to be synthesized is called the *target cost* [1]. The naturalness of the synthesized speech is strongly influenced by the match between elements concatenated together. This is captured by the *concatenation cost.* By definition, the concatenation cost of two neighboring elements from the speech corpus is zero, as the cut speech can be restored in its original natural form.

To investigate correspondence and matching, features are specified at the levels of phoneme, syllable, word and prosodic unit (such as a clause). Acoustic features of speech (such as the pitch and formant structure) are not currently utilized in our system, as we suppose that the prosodic features (such as the tone and modality of the sentence) hold sufficiently strong discriminative power. After tuning the weights of the cost function factors, the annotated speech corpus allows determining both the correspondence between a portion of speech and any part of the speech corpus, and identifying the fit between any two elements selected for concatenation.

Cost function factor weights can be adjusted by going through multiple iterations of listening test and modification phases. The correspondence of phonemes is not an absolute requirement, which has the important implication that phonemes of the same class can substitute one another, assuming that the concatenation cost is significantly decreased by this exchange. The utility of such a solution is explained by the fact that the imprecise phoneme may go entirely unnoticed by the listener if it fits well in the auditory environment (for example in an unaccented case).

Element selection cannot be done one by one because the fit of elements to each other must be taken into account. The goal of maximizing the overall quality of the produced speech makes a method similar to the Viterbi-algorithm [6] a plausible choice. Acoustic and prosodic effects overarching sentence boundaries can be disregarded, and therefore the target of optimal synthesis is a single sentence. The cost to be minimized is the sum of the target and concatenation costs for the entire sentence, over all possible selections of units.

2.3. Specifying element size

The peculiarity of corpus based synthesis is that in addition to making an element choice decision, the length of the element to be inserted can also vary [6,7]. When – in accordance with the requirements outlined in the previous section – the concatenation cost is zero for two adjacent elements from the speech corpus (the elements occurred together in the recording), then minimizing the cost function implicitly determines the size of the element as well.

However, this approach is not applicable in a real setting. The speech synthesizer is designed for a limited, but not closed domain, which means that knowing the target domain doesn't exclude the occurrence of new words (such as region names). To allow the synthesis of any arbitrary word, the system must be able to create speech from basic building blocks; diphone or triphone based synthesis must be available. The building blocks are necessarily the basic elements (diphones or triphones), if element size is specified with the help of a cost function. Search space can contain several million elements in this case, resulting in a long time to find the appropriate element – and eventually resulting in slow synthesis.

A possible solution is the *acoustic clustering* (AC, [8]) of elements, such that elements clustered together have minimal distances given by target cost function. Clustering can be done offline when annotating the speech corpus. The clusters can be used to reduce search space during synthesis. This approach has the advantage of not explicitly binding clusters to certain features.

In a different approach, longer elements (such as phrases, words or syllables) are also labeled in the speech database, and can be selected directly (without the implicit selection mechanism of a cost function)

[9,10]. The *Phonological Structure Matching* (PSM, [8]) algorithm first searches for an element to be inserted among the longer elements of higher levels. If this does not succeed, search is continued at a lower level. In the worst case (like when synthesizing a new word), the building blocks will be the diphone or triphone elements of the lowest level.

The PSM implemented in this manner still has to face the large number of different elements at the lowest level. This lead us to use acoustic clustering (AC) of diphones in our system below the segment level, while letting PSM select the element size above this level [8,11].

Consequently, at least one instance of all possible diphones must exist in the recorded speech corpus. To ensure this, we split the texts for the announcer to two parts, designed along different lines. The first part provides coverage of frequent words and phrases as determined from the statistical properties of the target domain, and allows selecting the longest possible elements for concatenation. The other part ensures coverage of diphones for the diphone based synthesis.

2.4. Database design and statistical analysis

The quality of corpus based synthesizers is fundamentally influenced by the construction of the speech corpus, which the element selection algorithm can later use to retrieve elements of varying sizes [8,12]. An efficient element selection algorithm assumes a well structured data storage solution. Care must also be taken to allow later potential extensions in the designed and realized database, without risking inconsistency.

The design of a well utilized speech corpus requires determining an optimal set of elements for storage in the database. Optimality in this case means finding an equilibrium between a large number of elements demanded by quality requirements, and a minimal element number constrained by performance considerations.

To help in finding an element set of optimal size and composition along these guidelines, we conducted some statistical analyses. Our investigations are based on a continuously growing collection of texts containing weather reports from various Hungarian sources on the Internet. The analysis database stores word forms and word form pairs, allowing statistical analyses for word forms, word form pairs and general statistical properties (such as the number and modality of sentences). A syllable-level analysis database is also under development.

The main table of the database contains word instances. Every word instance has an identifier and type (word, number, abbreviation, sign, punctuation), and the identifier of the preceding and following words, position in sentence, and sentence position in text are also stored. The position of the word in its sentence is recorded by two – numerical and structural – properties. The former means the number of the word in the sentence, while the latter shows whether the words is at the beginning or end of a sentence, preceding or following a comma, or in an enumeration. Any word instance can belong to more than one of these categories.

Before starting with the statistical analyses, we created a data table of abbreviations and their resolutions, storing the frequency of occurrence as well. A list of common misspelled words was also created, giving the correct form of the words and the frequency of the misspelled version. The construction of these tables was helped by a certain level of automation, but was mostly done by hand. In practice, abbreviation resolution was done by looking for features of words indicating an abbreviation, – such as three-letter long words containing only consonants are typically abbreviations, – and then reviewing the list manually. We used an external spell checking solution to collect misspelled words from our text corpus.

Our investigations revealed that the most common types of errors were mistakes in accentuation. The normalization of the text corpus was accomplished using the abbreviation and misspelled word tables. It is worthwhile to note that these tables will be of further use in automating the correction of new weather report texts.

The 20 sources of weather reports (such as *http://www.met.hu*) provided 56,000 sentences, containing 670,000 elements (words, numbers, abbreviations, signs and punctuations – signs are the “+”, the “-”, the “plus” and “minus” words) between April 2004 and May 2005. Some 493,000 of these are words (5200 distinct word forms), 43,000 are numbers, and the rest are punctuation and signs. Almost all sentences are statements; there were only a few questions and exclamations. On average, there are 10 words in a sentence (including numbers as well). The average length of words is slightly over 6 letters, which might seem unintuitive, as the list of most frequent words are topped by definite pronouns, one or two letters in length. The explanation lies in the frequent presence of longer than average weather related expressions (such as “hőmérséklet”, “várható”, “csúcsértéke”, “felhőzet”; *“temperature”, “expected”, “peak”, “clouds”*). The length of the longest word is 23 letters (“hőmérséklet-csökkenéssel”; “with decrease in temperature”). Hyphenated words were regarded as a single word (such as “Dél-Dunántúl”; *“South-Transdanubia”*).

A table of word length distribution was created, which showed that words between lengths of 6 and 10 appear in the most various forms. These, and further investigations of words include words in the traditional sense only, and do not include numbers and punctuation.

The frequency of word forms was also investigated: our list of words was labeled with the coverage percentage provided by each word. For the k^{th} word this means that a list of the k most frequent word forms would cover a portion of our weather report corpus; the size of this portion is given by the sum of coverage percentages for the k words.

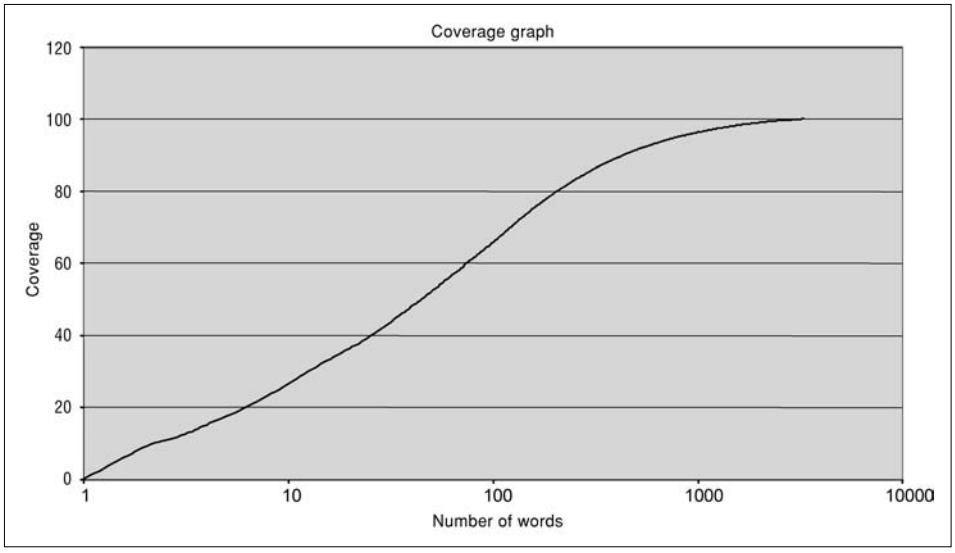


Figure 1. Coverage graph for weather reports

Our analysis led to the conclusion that the 10 most frequent word forms cover 31% of the input corpus. As little as 500 words ensure 92% coverage, while with 2300 words this reaches 99%. A corpus from an unrestricted domain requires approximately 70,000 word forms to reach 90% coverage [13].

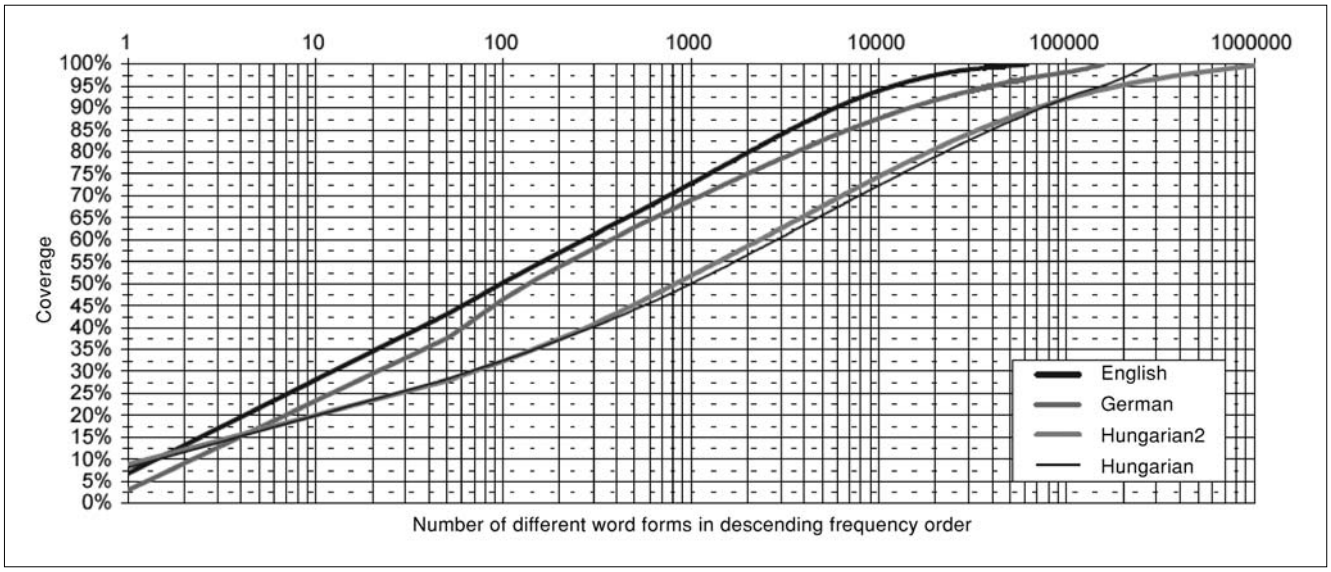
This favorable property is the result of the restricted domain, and in part a result of the limited annual coverage of the weather report corpus (the unavailability of archived weather report texts limited us to one year of our data collection period). Our further inquiries will concentrate on factoring out the latter. However it is important to notice, that these properties are almost exactly the same as the ones we got from our previous analysis based on a half year collection period, so we assume that these results are mainly due to the restricted domain. The coverage diagrams for the restricted and general domains are shown on *Figure 1.* and *Figure 2.* for reference.

Any word form must be available in multiple surroundings, as the pronunciation varies with the word's position in the sentence or with the different adjacent sounds. Considering this need, the final corpus will contain more word forms than the minimum required for simple coverage. This has led us to create statistics concentrating on word form frequencies in view of the left and right context. Frequent words (such as "hőmérséklet"; *"temperature"*) need to be recorded occurring in most of their frequent context types.

In order to take into consideration the position of a word inside the sentence, our data table was created to store information on this aspect as well. The actual position number is of little importance in a practical solution, since structural information captures location dependent word realization in sufficient detail. This structural information represents different pronunciation requirements as the beginning or end of a sentence, and before or after a comma (at clause boundaries and in enumerations). The design of the speech corpus has to incorporate this knowledge. In addition to investigating properties of individual words, the statistical analyses have to deal with words pairs (and word chains in the general case), to allow high quality synthesis of frequent phrases.

Finally, we created a list of foreign words appearing in weather reports (such as "Dubrovnik"), as these have to be synthesized differently. A possible solution is an exception dictionary, containing the correct Hungarian pronunciation of these foreign words ("Dubrovnik"

Figure 2. Coverage graphs for the general domain (source: [13])



would be translated to “dubrovnyik”, which corresponds to the Hungarian pronunciation of this word).

Considering the construction of the speech corpus, the design process has to select a small set from the available large text corpus, providing a good coverage for the entire corpus. A greedy selection algorithm is commonly used to achieve this goal [12]. This is a simple iterative solution, selecting entire sentences from the large text corpus, guided by a target of high coverage for the input corpus. Each iteration adds a sentence containing the highest number of words that are not yet covered by the selected corpus. An element is considered to be not covered if the corpus under construction does not contain an element with the same feature vector (a vector with parameters of interest (such as tone or intensity) as dimensions. Values along dimensions express how much a certain property is valid for the element. The iterative process terminates when the corpus under construction fulfills some predefined requirements (such as providing a given coverage ratio).

A key point to the success of the algorithm is the size and composition of the feature vector. A long vector results in high differentiation between elements, and results in the algorithm failing to cover most of the elements. A short vector will results in multiple coverage of most elements, prohibiting the selection of the most appropriate one. An optimal solution is not known, but several proposals outlining the composition of the feature space are available (such as emphasizing stress or the left and right contexts). A traditional solution associates a boolean value to features (criteria fulfilled or not), but fuzzy implementations, allowing a continuum of values between 0 and 1 also exist, such as in considering a match of left contexts. Stress is best represented with two possible values (stressed or unstressed). The success of the fuzzy solution mostly depends on the mapping of feature fulfillment levels to non-binary values.

The algorithm is generally used to obtain full diphone coverage, but can be extended to also select sentences that contain elements (words, word pairs) that are worthwhile for inclusion in the created corpus, based on their frequent occurrence shown by the statistical analyses. As mentioned previously, the investigations need to consider factors other than the frequency of occurrence, such as the context.

Database design must also consider the LNRE phenomenon (*Large Number of Rare Events,* [14]), which means that while the majority of elements rarely occur in speech, – and therefore each of these elements are rarely used in synthesis, – the number of such elements is so significant that at least one will be necessary for the synthesis of any given sentence.

Since the creation of the speech corpus focuses on including the most frequent syllables, words, phrases and sentences, the existence of the LNRE phenomenon implies that virtually all sentences requested for synthesis will contain portions that have no corresponding elements in the database. This means that the corpus must be constructed to include all possible diphones in at least one version, but the more frequent ones in multiple contexts. Such a design ensures that all portions of an input text can be synthesized to speech; using diphones in the worst case [11]. The number of possible diphones equals the square of the number of phonemes plus one (as silence can also be part of a diphone). However, not all diphones are necessary, as full diphone coverage can be achieved with at most a couple of thousand elements in European languages [15].

Once the speech corpus has been created along the above guidelines, it must be stored in a data structure allowing efficient element selection. This data structure is comprised of three fundamentally different parts.

The first is the collection of files containing the waveforms. Each file contains the annotated waveform of a single sentence. This solution, while ensuring a small file size, also allows using one file to load an element, since synthesis doesn’t require elements overlapping sentence boundaries.

The second part of the data structure contains the diphones. The directory of diphones is stored in a tree data structure, containing feature vectors and references to the files containing the diphones. The tree contains elements in the order of their appearance in the corpus, such that an inorder traversal of the tree returns the original corpus [16]. This is relevant in allowing a simple implementation of the varying element size selection. To select an element longer than a single diphone, the diphones returned along the inorder traversal of the tree initiated at the starting element can simply be concatenated.

The third part of our data structure speeds up search in the tree. We created a word tree [17], storing diphones such that nodes of this tree correspond to possible prefixes of the diphones, and the leaves of the tree contain references to diphones in the first tree.

When retrieving an element starting with a certain diphone, the leaf referencing this diphone is first located in the word tree (multiple matches are stored in a chained list). The requested longer element can then be obtained by the inorder walk in the corpus mapping tree.

The hierarchy described above can be improved by creating word trees for words, word pairs and sentences, in addition to making one for diphones. Creation of the data structure must ensure easy maintainability of consistency. As an extension of the database would take place before synthesis, allowing an update of the related search structures in addition to the update of the waveform files.

2.5. Listening experiments

We performed a proof of concept experiment based on the weather forecasts collected from the online archives of the Hungarian Radio. The weather forecasts of two consecutive days on radio stations Kossuth, Pe-

tőfi and Bartók were used. Although the linguistic content of these utterances were given, we could use them to evaluate the capabilities of the system being developed. Our aim was to predict the naturalness of the speech produced by a corpus-based system by manually synthesizing sentences of weather forecasts.

We analysed 149 weather forecasts by 22 announcers. There were just a small number of recordings available from each announcer that made the synthesis task harder than that of the system under development. Several words occur only once for some speakers. On the other hand, a number of words occur in almost every forecast since the recordings are taken from a two day period. During this short period there was not enough time for the forecasts to change much.

In order to get comparable results, we hand synthesized sentences on some announcers' voice that are also available on another voice. This is the rationale behind not choosing arbitrary sentences to be synthesized. The small corpus size was another reason. First, we selected the announcers with the highest representation in the corpus. Then we transcribed the forecasts belonging to them. In the transcript of the resulting 54 weather forecasts, we picked five sentences that consist of words that are also available by another speaker in a similar context. The details of the selected sentences are summarized in *Table 1.* where a horizontal line in a sentence denotes a concatenation point.

The stimuli for our listening experiment consisted of the five hand synthesized sentences, five natural utterances and five-five sentences synthesized by the diphone-based and the triphone-based Profivox text-to-speech system with automatic prosody (the triphone-based system is under development).

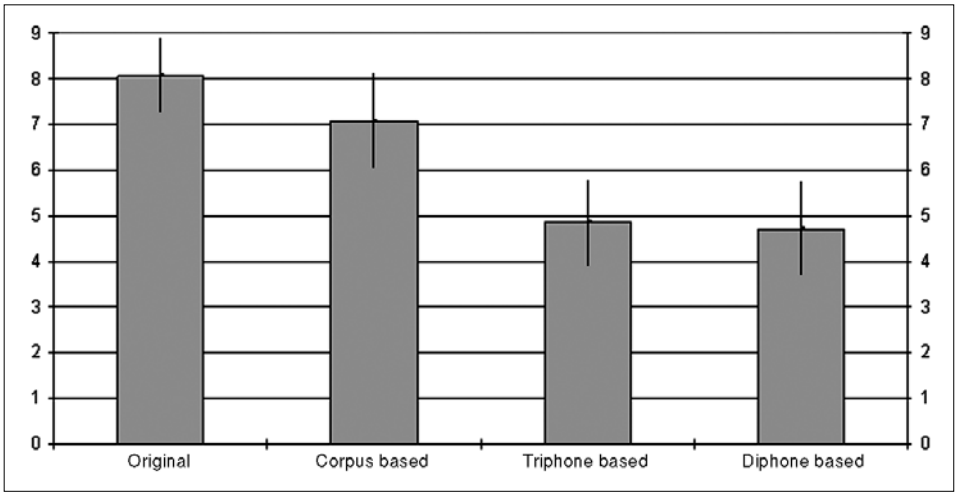


Figure 3.
Results of listening tests: average and standard deviation

The five native Hungarian subjects were instructed to rate the naturalness of each one of the 20 recordings on a scale between 1 and 9. The means and standard deviations of the ratings are shown on *Figure 3.*

According to the results, the triphone-based approach showed little improvement over the diphone-based system. Note that the former one is not a deployed system yet (among others, new signal processing algorithms are being implemented). The manually synthesized corpus-based sentences achieved a score two points higher than the systems employing concatenation of short, fixed-length units. The corpus-based stimuli was outperformed only by the natural utterances. The standard deviation is relatively small for each group that allows us to draw general conclusions. As it was expected, the scores of the natural utterances showed the lowest variability.

Table 1. Properties of synthesized sentences

Sentence	Original speaker	Speaker of synthesized sentence	# of elements concatenated
A hőmérséklet hajnalban mínusz | egy, | mínusz| hat, | holnap | napközben | mínusz | egy, | plusz négy fok | között alakul.	András	Adrienn	10
Napközben | országszerte várható csapadék, északon, | északnyugaton | havazás, | délkeleten eső, | másutt | havas | eső, | ónos | eső.	Erika	Klára	10
Végül az időjárásról: | mindenütt beborul | az ég, | reggelig egyre többfelé | lehet gyenge havazás, | hószállingózás.	Zsuzsa	András	6
A hőmérséklet | kora délután | kettő | és | hét | fok | között | alakul.	Zsuzsa	András	8
A nyugati, | északnyugati | szelet | sokfelé erős, | a Dunántúlon | helyenként | viharos | lökések | kísérik	Zsuzsa	István	9

Our experiment predicts a significant improvement in naturalness for the corpus-based text-to-speech system. Essentially, it is not expected that systems based on the concatenation of fixed-length units can achieve a comparable speech quality improvement. The corpus-based approach also has its drawbacks, such as a large database has to be recorded, labeled, stored and searched. Furthermore, the domain is limited and the computational complexity is far higher than for fixed-length concatenation.

3. Conclusion

The corpus-based approach to text-to-speech synthesis is a novel concept that has not been applied to Hungarian yet. It opens the way for synthetic speech to approach the quality of natural speech.

In this paper we outlined the fundamentals of this concept and gave a detailed progress report on the ongoing research at the Budapest University of Technology and Economics, TMIT Laboratory of Speech Technology (BME TMIT Beszédkutatási Laboratórium). The goal of this research project is to develop a limited domain, corpus-based TTS for Hungarian.

We gave an account of the statistical analysis performed on weather forecasts, the method of voice selection and discussed some other design issues. We conducted a listening experiment in order to predict the quality improvement achievable by the corpus-based approach.

Encouraged by the promising results, the next phase of our project is the implementation of the system. We are developing an algorithm for unit selection on multiple levels based on the results of our statistical analysis of the target domain. In the first version, we plan to define word and word N-gram levels and to restrict the input to a limited vocabulary. The second version would enable the synthesis of arbitrary words by backing off to a set of diphones based on acoustic clustering. The weights for the features used to calculate concatenation and target costs will be estimated by a sequence of iterative listening tests and refinements.

Acknowledgement

The authors would like to thank their colleagues at the BME TMIT Laboratory of Speech Technology for their invaluable help. We especially thank the Hungarian Radio for authorizing access to high quality weather report recordings.

References

[1] Bernd Möbius, “Corpus-Based Speech Synthesis: Methods and Challenges”, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), pp.87–116., 2000.

[2] Yi, J.R.W., Glass, J.R., “Natural-Sounding Speech Synthesis using Variable-Length Units”, Proc. ICSLP-98, Sydney Australia, Vol. 4, pp.1167–1170., 1998.

[3] Olaszy, G., Németh G., Olaszi, P., Kiss, G., Gordos, G., “PROFIVOX – A Hungarian Professional TTS System for Telecommunications Applications”, International Journal of Speech Technology, Vol. 3, Numbers 3/4, pp.201–216., Dec. 2000.

[4] Olaszi Péter, “Magyar nyelvű beszéd-szöveg átalakítás: nyelvi modellek, algoritmusok és megvalósításuk” (Hungarian Text-To-Speech Synthesis: Linguistic Models, Algorithms and their Implementation) PhD dissertation, BME, Budapest, pp.5–15., 2002.

[5] G. Olaszy, G. Németh, “IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method”, in D. Gardner-Bonneau ed., Human Factors and Interactive Voice Response Systems, Kluwer, pp.237–255., 1999.

[6] Jon Rong-Wei Yi, “Natural-Sounding Speech Synthesis Using Variable-Length Units”, Master of Engineering Thesis, Massachusetts Institute of Technology, 1997.

[7] S. P. Kishore and Alan W. Black, “Unit Size in Unit Selection Speech Synthesis”, Eurospeech 2003, pp.1317–1320., 2003.

[8] Antje Schweitzer, Norbert Braunschweiler, Tanja Klankert, Bernd Möbius, Bettina Sauberlich, “Restricted Unlimited Domain Synthesis”, Eurospeech 2003, pp.1321–1324., 2003.

[9] Eric Lewis and Mark Tatham, “Word and Syllable Concatenation in Text-to-Speech Synthesis”, Eurospeech 2001, Vol. 2, pp.615–618., 1999.

[10] Eric Lewis and Mark Tatham, “Automatic Segmentation of Recorded Speech into Syllables for Speech Synthesis”, Eurospeech 2001, pp.1703–1706., 2001.

[11] Michael Pucher, Friedrich Neubarth, Erhard Rank, Georg Niklfeld, Qi Guan, “Combining Non-uniform Unit Selection with Diphone Based Synthesis”, Eurospeech 2003, pp.1329–1332., 2003.

[12] Baris Bozkurt, Ozlem Ozturk, Thierry Dutoit, “Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection”, Eurospeech 2003, pp.277–280., 2003.

[13] G. Németh, Cs. Zainkó, “Word Unit Based Multilingual Comparative Analysis of Text Corpora”, Eurospeech 2001, pp.2035–2038., 2001.

[14] Ove Andersen, Charles Hoequist, “Keeping Rare Events Rare”, Eurospeech 2003, Vol. 2., pp.1337–1340., 2003.

[15] Dr. Gordos Géza, Takács György, “Digitális beszédfeldolgozás” (Digital Speech Processing, in Hungarian), Műszaki Könyvkiadó, pp.191–197, 1983.

[16] Rónyai L., Ivanyos G., Szabó R., “Algoritmusok” (Algorithms, in Hungarian), Typotex, p.60., 1999.

[17] Knuth, D. E., “A számítógép-programozás művészete”, (The Art of Computer Programming, in Hungarian), Műszaki Könyvkiadó, Budapest, p.503., 1988.