Towards More Intelligent Speech Interface

Gregor Rozinaj
Dept. of Telecommunications
Slovak University of Technology
Bratislava, Slovakia
Gregor.Rozinaj@stuba.sk

Abstract— The demand on information increases continuously. The use of computers has been changed from application and program development more and more to the communications and data retrieval. Multimedia interface of computers allows on-line collaboration not only in an office but even in extreme situations like driving, sports, entertainment activities etc. One of the principal modality in multimedia communications is speech. Recent research in speech recognition are devotes to continuous speech recognition. Basic algorithms for "word spotting" are slightly substituted by more sophisticated "topic spotting" methods. In the opposite direction of communications from computer to the user, the speech synthesis seems to be an easier task than speech recognition. However, naturalness of the artificial speech is a complex problem. An enormous effort has been made on speech prosody. Sinusoidal and noise models of speech seem to be a promising topic for artificial speech manipulation.

Keywords-mulimodal interface, speech synthesis, speech quality.

I. Introduction

The field of multimodal interfaces, computer graphics and speech recognition draw interest of researchers for a long time. This fact led to a development of increasingly more sophisticated autonomous or semi-autonomous virtual human interfaces over the last few years. Stephandis (C. Stephanidis, G. Salvendy, et all., 1998) have predicted that public information systems, terminals and information appliances will be increasingly used in a variety of domains.

The multimodal communications or multimedia user interface consists of a combination of basic modalities like image or video, audio or speech and data used for the communications with computer. The aim of this paper is to focus on one part of audio modality which is a speech synthesis. Speech synthesis means creating human-like speech using a machine, which is known as speech synthesizer.

There are several types of speech synthesizers, but each is made to do the same: to reproduce the given text in the clearest and most understandable manner.

Several basic approaches like synthesis using units concatenation, formant synthesis, articulation synthesis, HMM synthesis etc. are used for synthesis of human speech. Units concatenation method based on uniform or non uniform units is considered as the most natural way of speech synthesis in the present.

II. Speech Synthesis

Under the term naturalness of the synthesized speech we understand the similarity to the human speech in general. The speech is considered to be natural, if we are not able to distinguish in the context, if the source of the speech is a computer or a human. One of the main weaknesses of the speech naturalness in the process of synthesis is strong parameterization (e.g. 3 formants only) based on empirical settings during some synthesis algorithms (rule-based). This is the reason, that units concatenation methods offer the best naturalness of the synthesized speech. The quality of the synthesis depends in this case not only on the number of concatenations but on the character of these conjunctions, as well. The borders of each segment are slightly deformed according the context. Blind concatenation which does not respect this fact results in a non-natural prosody [1] [2].

The problem of quality synthesis is much more complex, if we take in an account correct pronunciation of each word, interpreting abbreviations, reading numbers and special symbols. This problem becomes much more difficult in flective languages, where the correct form of a word depends on a grammatical form coming from the semantic context. Last but not least, correct prosody of the speech leads not only to the naturalness, but intelligibility, too.

A. *Intelligent feedback for phonetic transcription*

People generally read the text correctly pronounced. This process we try to simulate on systems of phonetic transcription. In practice, this can be achieved by several methods. The first can be defined as knowledge transcript (knowledge-based) by LTS (letter - to - sound) rules, where the production rules are created with the help of phonetics - experts to establish generally applicable rules of LTS.

The second method is called the corpus method (data - driven) that automatically generates LTS rules under contextual phenomena found in the manualy described corpus. Set up rules are then applied to the input text.

The proposal is the creation of the transcription model that has to learn the rules of the phoneme pronunciation in a set of training sentences. The corpus method in a contrary to the knowledge-based method has one indisputable advantage that the transcription module is not dependent on an expert in phonetics.

Another method is the transcription according to the dictionary. The dictionary contains stored exceptions to the pronunciation of the specific language and phonetic transcription in the first step checks the dictionary. If the word is there, the phonetic transcription of his dictionary is used.

The disadvantage of all the above methods is that no one will not cover all the exceptions that may occur in the language. Such exceptions arise e.g. in domestification of words from foreign languages and their inclusion in the daily routine communications (eg the word Internet). The task remains to solve, how to make full and perfect transcript in speech synthesizer. In practice, this means to design the system that we can gradually modify or improve, respectively. In other words, the system is able to learn. That means a fusion system, which allows taking "online" modification of transcription, which will be stored somehow in the memory of the synthesizer.

B. *Learn the transcription*

We have already developed several possibilities for intelligent feedback or teaching the synthesizer online a correct pronunciation.

The possibilities are following:

Correction of the phonetic transcription in SAMPA alphabet per hand. Synthesizer interface allows the user to manually override erroneous word. After transcription, the user may again have to hear a synthesized text with the corrected word (or words).

The second variant is based on a selection of the word with incorrect transcription. The synthesizer offers the set of alternative pronunciations. Synthesizer interface provides the user (eg by clicking on the wrong word) all the ways it can be in the specific language pronounced. The options were drawn up under the rules of pronunciation. The set of pronunciation rules have been defined for Slovak as a specific language.. These rules were established by the mentioned corpus, the success rate of the pronunciation is around 97% [67]. On the basis of established rules, we can create different possibilities of pronunciation phonemes and offer the user a list of possible different pronunciations of the word repaired. Users select from the options offered to the right and may again have to hear a synthesized text with the corrected word (or words). If none of the options offered is correct, it is possible to manually repair the text according the previous method. Based on the manual corrections of the text it can be extended an alternative pronunciation of individual phonemes.

The third possibility is similar to the previous one but is more intelligent. All options from the previous paragraph should not be arranged randomly or alphabetically, but due to the probability and a certain logic. Synthesizer interface provides the user (eg by clicking on the wrong word) all options with the transcription where the incorrect transcription is gradually replaced by the possible vowels of the alphabet. Options are sorted logically with the highest probability at the beginning.

Last variant is based on acoustic correction of the pronunciation. This method of repair requires a perfect speech recognizer. User after hearing the synthesized text corrects the misspelled word or words using acoustic feedback, e.g. saying the correct pronunciation. He can also read the whole sentence in which the misspelled word or words were found. Speech recognizer recognizes the correct pronunciation/transcription of words written and the synthesizer can distinguish, which words have been misspelled and replace them. The user can again hear a synthesized text already corrected and the process may be repeated.

The whole process of repair phonetic transcription would not be very useful if the synthesizer would not learn the new correct pronunciation. To avoid correcting the same words again, the synthesizer should remember the "learned" new spelling words. It is necessary to find the optimal way of learning.

At the start of transcription we have the dictionary of words with transcription and a set of transcription rules. If there is a correction of the word, it is clear that the word is not in the dictionary and that the transcript has been driven under the rules. Thus, the correction word will be stored at the beginning of the list of exceptions. The list should include the original word, the corrected word, and the date of correction. Collecting and storing the words in the dictionary of exceptions, however, will result in enormous database after longer period of training and the whole process could be significantly slower. Therefore, after a reasonable number of records it is necessary to edit the rules according the obtained knowledge from the list of exceptions. If we found the sequence of vowels in the dictionary more times and always with the same pronunciation, we can make a new rule and set of rules will be extended. It is very important to find a threshold to determine whether a sequence of consonants (or word) will remain in the dictionary of exceptions or by the rules. So if there are several words with the same basis, it is preferable to make these words a new rule. However, if the new rule meant that the transcription algorithm has to face a word, it is preferable to retain the word in the dictionary. The learning process of transcription can be summarized in two basic points:

- storage of exceptions to the dictionary
- statistical evaluation of changes in pronunciation in terms of time and rules

III. Reading Abbreviations

Another big task for the speech synthesizer is reading abbreviations. Solving this problem will make synthesized speech more natural. Reading of abbreviations may be individual for each human, so this makes this task even more difficult. There are more possibilities how to read abbreviations: they can be replaced by words, spelled or pronounced if it’s possible.

It also important to consider how many time is the abbreviation placed and used in a text. If it is in article more times, it can be replaced by whole words (full meaning of abbreviation) first time and the other times can be spelled, or can be read in case it is readable - can be read and then spelled.

The reading of abbreviation is based usually on an existing. Of course, this glossary is never complete. Even, some abbreviation may possess several meaning. In case synthesized text is longer than one sentence, it is useful to know what the context of the topic. The main purpose of this is to detect the right meaning of one abbreviation.

The process how to improve already existing method for reading abbreviation is following.

- Preprocessing phase finds abbreviation in synthesized text. In case of longer article is this analyzed and based on abbreviation count is decided how to read it. The glossary is checked. If the abbreviation is in glossary, it is replaced by the text. In case of multiple occurrences sophisticated method can be used, e.g. the first time we can read a full meaning by spelled version in next occurrences, or it can be mixed – spelled and read full meaning.

- Next possibility how to increase the spontaneousness of expansion of abbreviation is simply read them. Of course not every expression can be pronounced fluently, e.g. expressions like „FIFA“, „UNESCO“ can be comfortably pronounced. So one part of processing should be analyses if abbreviation is vocable. In most languages it means if it contains vowels (a, e, i, o, u) or syllabical consonants. In that case spelling can be replaced by reading and synthesizers can choose randomly method for reading.

- In case the abbreviation is not in glossary, it is checked if the word is vocable. In that case it is read or spelled. In case it is not vocable, synthesizer spelled it every time.

- For intelligent reading and intelligent synthesizer it is necessary not only to complete the abbreviation glossary. In ideal case synthesize process should determine all parameters for abbreviation like word class, gender, case and number. So process of reading abbreviation is not isolated, it depends also on correct text preprocessing and determination off all word parameters.

A. *Learn the abbreviation interpretation*

The one separate area how to improve speech synthesis and reading of abbreviation is process of learning for speech synthesizer:

- Customize speech synthesizer interface so user can easily manually retype meaning of abbreviation (in case the abbreviation in the glossary has different meaning or it is not in the glossary). User can retype it like a text or in SAMPA alphabet – it depends on the synthesizer interface. User can again listen to the text with fixed word or words. In case the abbreviation is completely new, it is saved in temporary glossary. In case user changed only format of words this is stored in temporary glossary for word classes exceptions. From this temporary glossary new rules for word classes' determination are created.

- The higher level is that user interface offers all possibilities for abbreviation – in case of multiple meaning of abbreviation all meanings are displayed and for each of them also all grammar forms. These possibilities are listed based on glossary and temporary glossary. User chooses a correct form and can listen to synthesized text again. In case any of offered types of abbreviation is correct, user still can manually retype it (the procedure above).

- Acoustic modification of abbreviation. This method requires the perfect speech recognizer. User is listening to the synthesized text and in case of mistakes he reads the word or the whole sentence. Recognizer recognizes this new word or sentence, it writes it and replaces wrong one in synthesized text. Synthesizer compares glossary and temporary glossary and if this abbreviation is not there, it is automatically added into the temporary glossary.

The most important part is to find out how and what to store in temporary glossary. We should decide if to store only abbreviation, or also word before and behind it, if there should be only one glossary for new abbreviations and another one for all forms of abbreviation or if it should be together in one glossary. The other question is when to create new rule and what should stay in the glossary.

IV. Prosody modification of synthesized text

In general, the synthesized speech is artificial and unnatural. Therefore, in addition to its own synthesis, we try to adjust the prosody (pitch, energy, tempo), so we are as close as possible to the natural spoken language.

Speech prosody deals with properties of speech such as rhythm, intonation, pitch. It describes all the acoustic properties of speech, whose domain is not a simple phonetic segment, but larger units consisting of several segments, such as whole sentences. Therefore, prosodic phenomena are called supra - segmental. The phenomena we perceive as stress, accent modification or different intonation, rhythm and volume.

Prosody modification is applied in several ways. The first is a change of the pitch frequency. The pitch or basic glottal frequency is different for male, female and child voices. In order to adopt as closely as possible to the desired voice, pitch frequency should be adjusted according to the original. Effect of frequency is particularly significant for the voiced parts of speech, which can be considered to a periodic nature as the carrier of the pitch frequency. Since the unvoiced parts are not periodic in nature, they cannot change their frequency. The basic frequency of the glottal speech can be changed by making the period shorter or longer in individual periods of voiced phonemes. Making period shorter results in higher pitch frequency and vice versa, as described in PSOLA algorithm. Change of the period length causes undesired change of the length of phonemes. It is therefore necessary to adequately change the frequency of phonemes by adding or removing a couple of periods The length of the new signal is inversely proportional to the change in the glottal fundamental frequency. The more the frequency increases, the shorter the signal will be because of the reduction periods. Period to be shortened or extended are selected from the center of phonemes, because the centre is not influenced by the border of two phonemes. Within the process of prosody modification, the tempo and pitch frequency seem to be as deeply closely related parameters and cannot be viewed separately. Modification of the length in order to set up the correct timing of phonemes will be influenced by the subsequent translation periods and the length could change significantly. The opposite process, namely the translation to be made first and then add or remove as many periods of adjustment utterance length, would alter the number of periods in the resulting signal and the translation made depreciated. This problem can be solved in multiple ways, among primary solutions is considered progressive iteration, in which the duration of phonemes is changed first and then carrying out the translation to set the desired frequency. According the deviation from the required values is the process either closed or adjustment process is again repeated. This solution is time consuming and not ideal. Other known methods such as linear and quadratic interpolation, dealt with the problem more appropriately, while maintaining reasonable complexity of calculating

Change of tempo. Phoneme length mismatch causes audible and measurable difference in particular between the termination. Although everyone has a different pace or tempo, and every speech utterance of the same person may have a different rate, the specific pace of modification is quite accurate. Since the voiced and unvoiced phonemes are distinct, we have to approach the rate of modification using other ways. Unvoiced phonemes do not have any characteristic features to be taken in account within tempo modification. An extension of unvoiced phonemes is therefore quite simple and just means repetition, or removal of phonemes corresponding to a given extension or shortening. Modification of voiced phoneme length is not as trivial problem as a modification of the length of unvoiced phonemes. Voiced phoneme has a periodic character, which marks the features that we need to change the length of termination to be considered. Voiced part of speech cannot be extended or shortened by simply selecting the appropriate number of samples and their repetition or removal. Voiced speech adjustments must be made in accordance with the periodic nature of the signal, any adjustments must therefore be performed with all periods and their multiples. Extension of the signal in this case means adding the number of periods pertaining to that extension. Similarly, shortening the signal must be made by pertaining to the removal of the number of periods. The choice of periods and the operations with them are subject of many rules.

Change through acoustic feedback. This presentation will again require a perfect speech recognizer, similar to the methods mentioned above. The text would be recorded and the re-synthesizer synthesized text with no parameters speaker. At the beginning of the process would be necessary to determine the fundamental frequency of glottal speaking, speech rate, energy. These parameters are then compared with those of the database (assuming that the synthesized database includes all the necessary information.) The parameters of the database are adjusted to the original parameters utterance and the synthesis may follow. Thus synthesized text will have all the characteristics of the speaker. The appropriate way to these arrangements appears to be sinusoidal models [5].

V. Conclusion

In this paper, problems of the speech synthesis, one of the dominant part of a multimodal interface, has been analyzed and discussed. The topic is more complex than a simple text to speech module. Many modules for complex analysis and transformation of the text are necessary in the whole process. The idea of the intelligent system for speech synthesis with the learning ability based on the several forms of the user feedback has been introduced. This approach results in a system designed according the specific needs of the user.

Acknowledgment

This work has been supported by the Grant Agency of Slovakia, VEGA 1/0718/09 and APVV LPP-078-06.

References

[1] Mertens, P., d'Alessandro, C.: Pitch contour stylization using a tonal perception model. Proc.of 13th Int. Congress of Phonetic Sciences. vol. 4,Stockholm, Sweden (1995) 228-231

[2] Fujisaki, H., Ohno, S.: Prosodic parameterization of spoken Japanese based on a model of the generation process of F0. Proc.of ICSLP'96. vol. 4,Philadelphia,USA, (1996)2439-2442

[3] Hirst, D., Espesser, R.: Automatic modelling of Fundamental Frequency Using a Quadratic Spline Function. Travaux de l'Institut de Phonétique d'Aix en Provence, vol.15 (1993) 75-85

[4] Talafová, R., Rozinaj, G., Vojtko, J.: Speech Synthesis in Mobile Phones. Proceedings of 49th Int. Symposium ELMAR-2007, Zadar, Croatia (2007)

[5] Turi Nagy, M., Rozinaj, G., Čepko, J.: Design of an HNM System for Prosodic Modification of Slovak Speech. Croatian Society Electronics in Marine, 2006. – (2005) 147-150

[6] Tokuda, K., Zen, H., Black, A.W.: An HMM-based speech synthesis system applied to English, Proc. of 2002 IEEE SSW, (2002)

[7] Black, A. W., Lenzo, K. A. Building Synthetic Voices, Language Technologies Institute, Carnegie Mellon University, Retrieved October 12, 2007, from http://festvox.org/bsv

[8] Gaudissart, V., Ferreira, S., Thillou, C. SYPOLE: Mobile Reading Assistant for Blind People, Proc. of the 9th SPECOM 2004. St. Petersburg, Russia. 2004.

[9] Dusan, S., Gadbois, G., J., Flanagan, J. Multimodal Interaction on PDA's Integrating Speech and Pen Inputs, EUROSPEECH 2003. pp. 2225-2228. Geneva, Switzerland. 2003.