Special Speech Synthesis for Social Network Websites

Csaba Zainkó, Tamás Gábor Csapó, and Géza Németh

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
{zainko,csapot,nemeth}@tmit.bme.hu

Abstract. This paper gives an overview of the design concepts and implementation of a Hungarian microblog reading system. Speech synthesis of such special text requires some special components. First, an efficient diacritic reconstruction algorithm was applied. The accuracy of a former dictionary-based method was improved by machine learning to handle ambiguous cases properly. Second, an unlimited domain text-to-speech synthesizer was applied with extensions for emotional and spontaneous styles. Chat or blog texts often contain "emoticons" which mark the emotional state of the user. Therefore, an expressive speech synthesis method was adapted to a corpus-based synthesizer. Four emotions were generated and evaluated in a listening test: neutral, happy, angry and sad. The results of the experiments showed that happy and sad emotions can be generated with this algorithm, with best accuracy for female voice.

Key words: diacritic restoration, emotional speech synthesis, microblog reading system, chat-to-speech

1 Introduction

This paper gives an overview of the design concepts and implementation steps of a Hungarian microblog text-to-speech reading system. Microblog websites (e.g. Twitter, `http://twitter.com`) and chat-like talking applications are very popular nowadays. In chat applications, where little talk is written, it is advantageous to use speech instead of always keeping track of the dialog. The user can do something else than looking at the screen, and he will still know what is being said in the chat channel. This scenario is mainly useful when messages do not arrive very often. A microblog reader system can also be useful in mobile environment, because there is no possibility to continuously watch the display or the user does not have a free hand to handle the device (e.g. during car driving, or sport activities like running). Another possible situation is if the user is working with a full screen desktop application and he needs real time information from social networks. Loud reading demands only short time attention, and task changing is not necessary. This system is very useful for visually impaired and blind people, as well.

However, chat-to-speech synthesis conveys some new problems. In current web-based social networks people tend to use the special form of their language

(e.g. letters without diacritics for Hungarian and with "emoticons"). Text repairing algorithms are needed to recover the proper text that can be read by a TTS. Spontaneous style and emotional synthesized speech can help to improve how people accept these systems. There exist several systems that are specialized in chat reading. For example, an attempt has been made to fit together the free Espeak utility with the X-Chat IRC client [1]. A couple of other applications exists which use a TTS to read any type of documents, including web pages and blog sites, but most of them are prepared without fitting a general TTS to the specific task of chat reading.

Our approach is a first step in developing a Hungarian microblog reading system with more complex functions. Section 2 introduces the problem of missing diacritics and a combined machine learning approach to solve it. Section 3 discusses several approaches to make synthesized speech more spontaneous and thus more human-like. Section 4 shows a method to transform neutral speech to emotional. The last section summarizes and concludes the paper.

2 Diacritics Restoration

Characters with diacritics occur in large numbers in most European languages. According to [2], only the English alphabet is without diacritics from 36 languages studied. In several telecommunication applications, like SMS in cell phones, some or all diacritics of the input text are removed because of character encoding. On many small devices the typing of diacritic letters is uncomfortable and slow so people tend to use the diacritic-less letters of their language when writing computer documents and Web 2.0 websites. This is a hard problem for a text-to-speech system: the errors in the spoken utterances are much more confusing than in written text. Therefore, we apply a diacritic restoration algorithm which can formulate proper input text for a speech synthesizer.

The Hungarian language has five ambiguous sets of letters from the viewpoint of diacritics. These include nine letters with a diacritic, all of which are vowels. Some of the ambiguous sets differ only in quantity ("i-í", "o-ó", "ö-ő", "u-ú", "ü-ű"), while the others may differ in quantity and vowel quality as well ("a-á", "e-é", "o/ó-ö/ő", "u/ú-ü/ű").

Diacritic restoration is a well-known problem, and there are several methods to solve this task. With dictionary-based solutions up to 90% accuracy was reported in accent restoration depending on the language [3]. The use of Hidden Markov Models can lead to almost perfect restoration performance [4]. Most word based methods need a reliable morphological analysis tool. On the other hand, letter based methods are much easier to build and provide generalization beyond words [2].

A former dictionary-based algorithm for Hungarian selected always the word variant with the most possible diacritic pattern [5]. This method gives poor results when the variants occur nearly equally often. As this solution does not have generalization capability, words that are not included in the trainig corpus cannot be handled properly.

2.1 Combined Diacritics Restoration Method

We use a combined method applying a word level dictionary and a letter level machine learning approach together. The unambiguous cases (words with only one possible diacritic pattern, e.g. *az=the* is a Hungarian article, but *áz* does not occur as a Hungarian word) are handled using a dictionary, while the diacritics of the ambiguous cases (words with more than one possible diacritic patterns, e.g. *meg-még=plus-still*) are calculated using a decision tree.

First, a learning phase is performed. As training data, the Hungarian National Corpus (HNC) of 187 million words is applied [6]. HNC is a collection of written linguistic data representing present-day standard Hungarian. There are 3.58 million different word forms in the corpus (880 thousand without any diacritic letter and 2.7 million with at least one diacritic letter).

The first step of the learning phase is the separation of ambiguous and unambiguous words from the viewpoint of diacritics. For this, we use a naive algorithm: those words are marked as unambiguous, which have only one diacritic pattern variant, and the rest build up the ambiguous set of words. Those words are included in the dictionary as well, which have one diacritic pattern in more than 95% of the cases. These unambiguous cases cover about 84.5% of the input text. To handle typos, we applied a spell checker only on the ambiguous part, because on the unambiguous part the spell checker threw out too many valuable correct forms. The ambiguous part (15.5% of the corpus) is handled with a J4.8 decision tree [7], which is the open-source and improved implementation of the popular C4.5 decision tree. This learning phase was conducted using the Rapidminer data mining program (`http://www.rapidminer.com`). The ambiguous set of the input data is about 29 million cases. The parameters of the J4.8 tree were optimized in order to get an acceptable-sized tree that can be handled in an application. In order to improve accuracy of the diacritic restoration, the 100 most frequent ambiguous words (covering 60 % of wrong decisions of the former method in the ambiguous set) were trained with separate decision trees for each. During the decision tree learning phase, the context of 20 letters of the ambiguous vowels is extracted as the training data. This is twice the window as suggested by [2], but we intended to treat correctly very long words as well.

After the learning phase, the missing diacritics of the input text can be determined using the above-mentioned algorithm. First, the input text is separated into unambiguous and ambiguous words. The diacritized versions of the unambiguous words are searched in the dictionary. The context of the ambiguous words is calculated from the input text. After that, the J4.8 decision tree is applied and the restored diacritics are given as the output text.

2.2 Accuracy of Diacritic Restoration

Word accuracies were calculated because for the TTS domain correct words are more important than reconstructed vowels alone. Partly incorrect diacritic reconstruction degrades the quality of speech.

With our combined method, we applied training and validation on the "DIA", "Personal" sub-corpora and the whole HNC database. Accuracies of 97.7% for "DIA", 97.2% for "Personal" and 98.2% for the whole HNC can be reached in diacritic restoration applied to the Hungarian language. Detailed sub-results for the ambiguous cases only are shown in Table 1.

Table 1. Results of diacritic restoration for ambiguous cases.

	Baseline amb.	Top100 amb.	Ambiguous
DIA (DLA)	75.9%	92.7%	82.4%
Personal	71.4%	88.5%	80.8%
HNC (all)	75.2%	92.9%	82.9%

The three rows are the different test sets. The first ("DIA") is the Digital Literature Academy which was used in [2] as well. The second ("Personal") is a collection of the web forum sub-corpus of HNC, which is the most similar in topic to our target application. The third is the whole HNC database. The "Baseline amb." column gives the results of diacritics restoration applied to ambiguous cases when the decision is always the most likely form, as in the algorithm introduced in [5]. The next column shows the results of the 100 separately trained words. The third column gives the results of the rest of words that are known as ambiguous and are not contained in the previous column.

The tendency of percentages is similar in all of the test sets. The most relevant test set is the "Personal" for us. The word accuracy increased from 71.4% to 88.5% with 100 separately trained words and to 80.8% with the other ambiguous words. The trained decision tree can restore correct diacritics in 70.2% of the general cases (not included in Table 1). This number is an estimate for the accuracy of the solution for the out of dictionary words.

3 Spontaneous Synthesized Speech

Spontaneous speech is the most natural oral expression of humans. However, speech synthesis has mainly focused on read speech (e.g. in the form of huge read corpora) because it is much easier to process than spontaneous speech. According to [8], it is advantageous to mimic some aspects of spontaneous speech in a TTS system. This style is particularly useful in our chat and microblog reading system. It should be chosen, which properties of spontaneous speech are worth to be modelled in a human-machine communication system. For example, hesitation and humming are significant attributes of everyday speech, but they increase the cognitive load of the listener, thus disturb the understanding of a speaking machine system.

3.1 Corpus-based Unit Selection TTS

The unit selection TTS that was used in our experiments is described in detail in [9]. The currently used speech databases contain sentences from several domains (e.g. weather forecasts and radio news). The synthesizer can generate the prosody in two ways, depending on the type of the input sentence. If the sentence fits in the domain of the corpus, a simple prosody model is used, based on the relative position of words within a prosodic phrase. Because it is based on words, it will work properly only if most words of the input sentence are found in the corpus. If the sentence is out of theme, there will not be enough whole words, which can determine the prosody. On those parts of the sentences a template-based F_0 generation method [10] is applied. The templates are based on spontaneous speech. The obtained F_0 values are used in the target cost function of the TTS to follow the F_0 curve. This extension can help to use the originally limited domain TTS in the unlimited domain of chat and microblog reading.

3.2 Spontaneous Style Synthesized Speech

Several spontaneous like synthesized speech examples are generated, in the form of modifying the output of the corpus-based speech synthesizer. These utterances included some properties of spontaneous style speech. First, filled pauses were added, in the form of breathing, after the conjunctive words. Second, silent pauses were lengthened to mimic "thinking" during everyday speech. Third, hesitation and humming were added randomly and the F_0 curve was shifted. At last, the structure of the sentences (e.g. word order) was modified in order to come closer to spontaneous speech.

3.3 Listening Test and Results

Two sentences were chosen for evaluation in a listening test. The variants of the sentences are shown in Table 2. The first variant was produced with a diphone system using copied natural prosody. Variants 2-5 were generated with the corpus-based unit selection TTS. The second variant contained the sentence with a re-edited structure. In the third variant, we applied the baseline position-based prosody of the TTS. In variants 4 and 5 hesitation, silent and filled pauses were added and the F_0 contour was modified to approximate spontaneous speech. The last variant was a natural spontaneous speech sample.

A small web-based listening test was conducted in order to get feedback from speech scientists. The goal of the test was to investigate, how people accept the modelled properties of spontaneous speech. After listening to each of the 6-6 variants of the two sentences, the listeners had to answer two 5-point MOS questions: Q1) "To what extent is this speech sample natural?" 5 - very human-like, ... 1 - very machine-like; Q2) "To what extent is this speech sample spontaneous?" 5 - totally spontaneous, ... 1 - absolutely not spontaneous.

Seven listeners evaluated the sentences (all of them were Hungarian phoneticians or speech synthesis experts; 3 male; 4 female; mean age: 33; mean test

duration: 4 minutes). The results are included in Table 2. According to the results, the insertion of breathing and the longer pauses helped to approximate the spontaneous speech, but the quality of speech was decreased. Despite of the frequent occurrence of hesitation in human speech, listeners of the test did not prefer the synthesized sentences with inserted hesitation. The diphone-based variant achieved very high spontaneous score, because its prosody was copied from the natural sample, but its quality is lower than that of unit selection TTS samples.

Table 2. Variants of speech samples used in the listening test and their MOS results.

			Sent. 1		Sent. 2	
Num	Technology	Type	Q1	Q2	Q1	Q2
1	Diphone	Rule based prosody	2.00	3.86	2.25	4.38
2	Corpus	Reedited sentence structure	2.50	3.00	2.38	3.00
3	Corpus	Position based prosody	2.13	2.38	2.13	2.75
4	Corpus	Added hesitation	2.13	3.38	2.88	3.50
5	Corpus	Added hesitation, pauses; modified F_0	2.63	2.88	2.88	3.50
6	Human	Natural	5.00	5.00	5.00	5.00

This simple approach showed that it is possible to model several properties of spontaneous speech in a TTS system. In a real application, users tend to accept only quiet breathing and some laughter. Several sentences should be read together in one prosodic phrase, as in spontaneous speech we also talk in longer units.

4 Emotional Synthesized Speech

Usually there are few coherent sentences in a microblog, therefore the reader or listener may misunderstand the content without emotional signs. During writing the bloggers use "emoticons", but these cannot be read in themselves. These emoticons modify the meaning of the previous word or sentence. During speech synthesis it is beneficial to modify the natural sentences to their proper emotional sentences. In order to be able to quickly identify the emotional type of the heard sentence we use four emotions: neutral style and angry :@, happy :), sad :(. The "Personal" subcorpus of HNC contains 90 thousand emoticons (in 1.5 million sentences).

4.1 Emotion Modification Algorithm

The emotion modification method was chosen considering many points of view. Bulut et. al. [11] emphasize that for creating strong emotions (like angry and happy), the segmental components of speech are important features. Prosody is also important, but if the segmental structure is set properly, the emotion will

be identifiable with a less correct prosody as well. It can be used in such speech synthesis technologies where prosody modification is limited or not allowed (e. g. the applied corpus based unit selection system). The modification method should be applied on the output synthesized speech signal because this corpus-based TTS does not encourage strong signal modification.

Přibilová and Přibil [12] describe a method which is based on spectrum modification. A non-linear frequency scale transformation is applied on the speech spectral envelope. The main suprasegmental features are also modified: F_0, energy and duration. During human emotional speech, individual formants are shifted. This is caused by the physiological change of the vocal tract, and the distribution of low and high-frequency energy is also changing. This is the concept of spectral transformation used in emotion modification.

Our spectral transformation method is based on the PSOLA algorithm and uses nonlinear frequency scaling as suggested by [12]. First, the pitch markers are determined by the Praat program or the synthesis system computes them. An asymmetric Hann window function was used. The center of the window is at the pitch marks and the left and right end of the window are at the previous and next pitch marks. This time domain windowed signal is converted to frequency domain with the DFT (Discrete Fourier Transform) algorithm. The amplitude part is modified with a transform function [12] and the low-frequency and high-frequency energy distribution is set properly. The modification of the phase spectrum is not necessary. With an inverse DFT algorithm we convert back the signal to time domain. Before finishing the speech output, a second Hann window corrects the left and right ends of the signal to remove discontinuities. The segments are recombined with F_0 and duration modifications similar to the original PSOLA algorithm.

4.2 Experiments

Several emotional variants of three sentences with a male and a female voice were tested in a listening experiment in order to verify our hypothesis that the spectrum modification can improve emotion transformation. One sentence was the modification of a natural speech sample uttered by a professional female speaker, while the other two sentences were the modifications of the output of the corpus based speech synthesizer with a female and a male voice. The modifications included the methods described in Sec 4.1.

The modification parameters were similar to [12]. γ_1 and γ_2 are varied in the sentences with ± 0.05. F_0 is increased by 15% for anger, 17% for happy and decreased by 16% for sadness. Energy is increased by 4.6 dB for anger, 2.3 dB for happy and decreased by 3dB for sadness. The spectral energy distribution is also modified. For sadness the low-frequency components are increasedby 4 dB and for the other two emotions the high-frequency components are increased by 2 dB for happy and 4 dB for angry. The cut-off frequency was 1 kHz.

A separate listening test was conducted to determine the perceived emotion, naturalness and quality of synthetic sentences. The test was web-based, in order to emulate the circumstances of a potential application. 25 native speakers of

Hungarian participated in the test with no known hearing loss. The results of 3 listeners were excluded from the evaluation because they either did not finish the test, or were found to respond randomly. The remaining 22 listeners consisted of 15 male and 7 female testers having a mean age of 38 years. 8 listeners used head- or earphones while 14 testers listened to loudspeakers. The listening test took 9.6 minutes to complete, on average.

The listeners had the option to replay a stimulus as many times as they wished, but they were not allowed to go back to a preceding stimulus, once they rated it. The playback order was randomized individually for each listener.

4.3 Results of the Listening Test

The confusion matrix of the planned and recognized emotions is shown in Table 3, for the three sentences separately. From the 3-3 parameter variants of the sentences, the 1-1 best were selected from the viewpoint of highest recognition of intended emotion. Only these are included in the table.

With the female and male voice TTS, the sad emotion could be produced with the highest accuracy. In the female case, the happy emotion is acceptable as well, while in the male case, listeners mostly misrecognized the happy variants. The angry emotion could be generated better with the male TTS. Emotion modification caused high accuracies in happy and sad natural speech, while the angry emotion was less successful. Přibilová and Přibil [12] report on similar results, the worst identification is for angry, and sadness gets the best scores.

Table 3. Confusion matrix results of the emotion recognition

		Recognized															
		N	A	H	S		N	A	H	S		N	A	H	S	
Planned	N	82%	0%	14%	5%	N	45%	23%	0%	27%	N	77%	0%	18%	5%	N=neutral
	A	27%	27%	41%	5%	A	36%	41%	0%	23%	A	45%	32%	14%	9%	A=angry
	H	27%	5%	68%	0%	H	41%	23%	14%	23%	H	9%	5%	82%	5%	H=happy
	S	9%	5%	0%	86%	S	14%	5%	0%	81%	S	23%	5%	5%	68%	S=sad
		TTS-female					TTS-male					Natural-female				

According to the comments of the listeners, the emotional modification decreased the quality of speech only in the angry case. They reported that the utterances were more natural for the female sentences. This can be explained by a note of [12]: the parameters of the applied spectral modification algorithm were worked out for female speech. This test showed that for male speech, parameters should be applied in some other range.

5 Summary

In this paper, the concepts of a Hungarian chat and microblog reading speech synthesis system were introduced. We extended a former dictionary-based dia-

critic restoration algorithm with machine-learning applied at letter level, getting a combined method. This approach can lead to 98.2% accuracy for general topic input text and to 97.2% accuracy on the specific text of forum entries.

A limited domain corpus-based text-to-speech synthesizer was extended for use in the specific task of chat reading. Several properties of spontaneous speech were investigated. Some of them were integrated in the TTS system in order to create "loose style" machine-generated speech that is closer to that of used in our everyday human-human communication. The output speech of the synthesizer was passed into a spectral modification algorithm in order to produce emotional speech. Four emotions were investigated, of which the female version of happy and sad could be synthesized in the best quality according to a listening test.

This paper is an exploratory study in developing a specialized speech synthesis system. The proposed method can be applied in a chat reading program or with any social network web-site (e.g. Twitter). Our future plans include the completion of such a system.

Acknowledgments. We thank the listeners for participating in the listening tests. This research was supported by the TELEAUTO (OM-00102/2007) project of the Hungarian National Office for Research and Technology and by the ETOCOM project (TÁMOP-4.2.2-08/1/KMR-2008-0007).

References

1. X-Chat Text-To-Speech, `https://launchpad.net/xctts`
2. Mihalcea, R., Nastase, V.: Letter Level Learning for Language Independent Diacritics Restoration, In: COLING 2002, Taipei, Taiwan, pp. 1–7 (2002)
3. Galicia-Haro, S. N., Bolshakov, I. A., Gelbukh, A. F.: A Simple Spanish Part of Speech Tagger for Detection and Correction of Accentuation Error, In: Proc. of TSD, Plzen, Czech Republic, pp. 219–222 (1999)
4. Simard, M.: Automatic Insertion of Accents in French Text, In: Proc. of Conf. EMNLP, Granada, Spain, pp. 27–35 (1998)
5. Németh, G., Zainkó, Cs., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kiss, P.: The design, implementation and operation of a Hungarian e-mail reader. Int. Journ. Of Speech Techn. Vol. 3-4: pp. 216–228 (2000)
6. Hungarian National Corpus, `http://corpus.nytud.hu/mnsz`
7. Witten, I. H., Frank, E.: Using the J4.8 Decision Tree, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann (2005)
8. Carlson, R., Gustafson, K., Strangert, E.: Modelling Hesitation for Synthesis of Spontaneous Speech, In: Proc. of Speech Prosody, Dresden, pp. 69–72 (2006)
9. Fék, M., Pesti, P., Németh, G., Zainkó, Cs., Olaszy, G.: Corpus-Based Unit Selection TTS for Hungarian. In: Proc. of TSD, Brno, pp. 367–374 (2006)
10. Csapó, T. G., Zainkó, Cs., Németh, G.: A Study of Prosodic Variability Methods in a Corpus-Based Unit Selection Text-To-Speech System, Infocommunications Journal, Vol. LXV, Budapest, pp. 32–37 (2010)
11. Bulut, M., Narayanan, S. S., Syrdal, A.K.: Expressive Speech Synthesis Using a Concatenative Synthesizer, In: Proc. ICSLP, pp. 1265–1268 (2002)
12. Přibilová, A., Přibil, J.: Spectrum Modification for Emotional Speech Synthesis, Multimodal Signals: Cognitive and Algorithmic Issues, pp. 232–241 (2009)