Multilinguality, singing synthesis, acoustic emoticons, and other extensions of the Slovak speech synthesizer for SMS reading

Milan Rusko, Sachia Darzagin, Marian Trnka

Department of Speech Analysis and Synthesis
Institute of Informatics of the Slovak Academy of Sciences
milan.rusko@savba.sk

Abstract

After the successful implementation of our speech synthesizer Kempelen in the SMS to voice service of Eurotel, one of the biggest mobile telephone operators in Slovakia, we have found out that several extensions need to be done to make the service even more important for users. In the paper we describe the Slovak speech synthesizer Kempelen, its Hungarian version, we discuss the need of male/female/children voices. We describe a program enabling developers to create a new voice for testing purposes in 10 minutes. We introduce a new feature – speech singing synthesis in SMS to Voice service. Some basic acoustic and signal processing problems of the singing voice synthesis using speech synthesizer are discussed. Last, but not least we introduce a classification of Acousticons (pre-recorded illustrative sounds) which offer much more possibilities, than only acoustical representation of graphical emoticons. The acoustic, aesthetic, psychological, sociological and commercial aspects were taken into account in the classification scheme design.

1. Introduction

Being a country with a relatively small market, Slovakia always has to wait for introduction of new technologies a little bit longer, than the bigger countries. This was the case also with introducing the service of delivering the SMS messages in the form of the synthesized speech, known as SMS to Voice. The mobile phone operator Eurotel decided to start the service in June 2003. The core of this system is realized by the concatenative diphone speech synthesizer in Slovak, KEMPELEN, developed at the Department of Speech Analysis and Synthesis of the Institute of Informatics of the Slovak Academy of Sciences.

2. SMS to Voice peculiar features

The SMS to Voice, which was originally developed mainly to deliver important SMS messages to the addressees in the fixed telephone network is in the real operation often used for sending jokes and entertainment, messages carrying emotional content, advertisement and many others. So we decided to enrich the possibilities of the synthesizer in its further version by adding the following functionalities:

- more voices – male, female and children’s voice in the basic configuration
- second language (Hungarian), which is the mother tongue of the 10% of inhabitants of Slovakia
- singing voice synthesis with an easy code for melody and text input
- acoustic representation of emoticons (smilies)
- significantly broadened set of the acoustic icons – illustrative sounds for possible enrichment of the SMS messages with an easy to remember code for their insertion

3. KEMPELEN speech synthesizer

Speech synthesizer program KEMPELEN is based on concatenation of small elements of a pre-recorded speech signal (representing combinations of two phonemes called diphones). We use our own algorithm of concatenation, similar to that of PSOLA [1.] The pronunciation is controlled by the block of orthographical-to-orthoepical conversion based on a sophisticated set of rules supplemented by a pronunciation vocabulary and a list of exceptions [2.].

4. Voices

It can be important for the user to choose between female and male voice. A young woman would probably prefer male voice to listen to words of love sent to her by her boy friend. So we have implemented a possibility to switch between male and female voice wherever in the SMS message. We are working on the child voice now. It takes several weeks to build a new professional quality voice. To get an idea how the new voice will sound, we have designed a program that interactively records a set of nonsense words uttered by the speaker and immediately after the 10 minutes long recording session it automatically finds the needed diphones in the signal and creates a database for a draft new voice. The timbre of the new voice is the same as it will be in the definitive version, only the appearance of concatenation discontinuities and rhythmical mistakes is higher.
5. Towards multilinguality

One of the main trends in telecommunication services, information systems and computer speech interfaces is multilinguality. Developing English, German or French version of our synthesizer would probably not give any sense, as there are lots of high-quality synthesizers available for these languages, developed by reputable companies which have incomparably better financial and personal conditions for there development. We have however decided to make a Hungarian version of our synthesizer to broaden the rank of possible users by the Hungarian speaking fellow-citizens. We have used our synthesis engine and with a help of the students of Hungarian nationality and the employees of the Department of Hungarian language of the Comenius university in Bratislava we have defined rules and designed a database of synthesis elements and block of pronunciation for Hungarian. As a result we have a synthesizer in which one can switch between the two languages.

As we think, that the source texts for element-database building in the two languages should be spoken by the same person to keep the homogeneity of the timbre, we have started experiments with bilingual speakers, speaking both languages perfectly, without any foreign accent.

Foreign words included in the list of exceptions could also be treated the same way i.e. uttered by the same speaker. A problem that occurs in this approach is well known from unit selection synthesizers - the same recording conditions and emphasis of the speaker can be hardly guaranteed after a longer time interval.

6. Singing voice synthesis

Singing voice synthesizers have in general different purpose than speech synthesizers and they work on different principals. They are designed to provide enjoyable singing voice where intelligibility is not of high importance. They are designed like a specialized systems which can employ principals of music samplers, advanced methods of pitch processing and time stretching algorithms etc.

We decided to use the simplest and cheapest way – that is „to force the speech synthesizer to sing“. The basic formula for tempered tuning is:

\[ f_{n+k} = kqf_n \]  

where \( q = 1.05946309 \), which is the twelfth root of two and \( k \) is the number of half-tones between \( f_n \) and \( f_{n+k} \).

It is obvious, that a direct mathematical representation of a note code does not give an acceptable pitch contour for the singing voice synthesis even in the telephone application. Our analyses of the pitch contours of recorded songs showed that at least several phenomena should be taken into account, such as rise and fall times of the tones, and vibrato, its depth, envelope and frequency. The introduction of these changes improved the synthesized singing significantly [3].

A correct division of words into syllables seems to be difficult task. There role of syllables in singing is different than that in speech. Unvoiced consonants do not carry pitch information and the onset of the tone carrier – the vowel – can be shifted in time too much if the length of the initial consonant is not predicted well or if the consonant stands at a wrong position. Mistakes in dividing words into syllables cause a serious deterioration of synthesis quality. But as the text is input in syllables in our system, the problem of correct division is handled by the user.

Dynamic prediction of phoneme lengths in respect to the musical rhythm is the most complicated problem that we had to solve in our synthesizer.

A drawback of using concatenative speech synthesizer and its speech derived samples (mostly diphones) for singing voice synthesis is a metallic/robotic sound of longer tones and unnatural timbre mainly in higher tones.

In spite of imperfection of the solution we consider our singing voice feature to be fully functional and suitable to enrich the SMS to Voice service.

6.1. Melody and text coding in singing voice synthesis

There are minimum three parameters that must be sent to the singing synthesizer - note pitch, note length and text of the syllable. After experiments with our own input code, we decided to use simplified form of Ringing Tones text transfer language (RTTTL) [4]. This solution makes it possible for the user to prepare the melody in any of the available ringing tone melody editors. At first the melody in this code is sent and then the text follows - divided into syllables.


7.1. Emoticons (smileys) and their acoustic representation

The use of the emoticons (smileys) has been popular mainly in the E-mail texts. Some of them are also used by senders of SMS messages. A modern SMS to Voice system must be able to understand these signs and must be capable of their acoustic interpretation. The solution is not difficult – it is enough to have several files with typical sounds representing different moods and feelings of a human being. The appearance of the :-) emoticon could for instance call a sound file with recorded
children giggling, and the emoticon : ( would start playing of a recorded sad sigh etc.

Anyway there is not a one to one relation between the information content of the graphical emoticon and a sound file. Both of them can express a wide variety of meanings, moods and feelings. So we decided to widen the set of sounds that could be included in the acoustic SMS.

Firstly we made an analysis which kind of sounds are attractive for the user and then we grouped the related sounds into classes and we made a proposal of an easy to remember code to insert the acoustic emoticons or illustrative sounds in the SMS message.

7.2. Extended set of the illustrative sounds – Acousticons

The first was the analysis of the requirements of the users. We summed them up into several points:
- sounds and work with them must be a fun
- sounds must be clearly recognizable and well known to the user
- sounds must be vivid and must contribute to the expressiveness of the message
- it is desirable that the sound can express also secondary meaning
- the set must contain sounds that express for instance
  a) feelings or instantaneous mood of the writer
  b) attitude to the information contained in the text
  c) greetings, common phrases and congratulations in an amusing form
  d) the sounds that can neutrally illustrate or enrich the text of the message (e.g. symbolically represent the situation, mood, occasion, event, feast, frame of mind, condition, position, weather
  e) expressively or humorously illustrate the situation

The sounds can be realized as acoustic displays of human being, either using the voice organs (speech, singing, sigh, cry, whistling...) or other way (snap, applause). They can be also acoustic displays of animals (rooster crow, cow’s moo, etc.), sounds of an environment (siren, gun shoot, car brakes, opening of champagne), or sounds played on the musical instruments (e.g. gong, or a fragment of Jingle Bells or marriage march).

In the systematization of the sounds we were inspired by the General MIDI specification, so we called our system General SOUNDI (General System of Sound Icons). Such a structure makes it possible for the operator to change certain sounds for new, better ones (with the same meaning, same content, but better quality) whenever needed. Accepting the definition of General SOUNDI by the synthesizer designers would lead to a compatibility of the illustrative sound sets. The user would then be able to utilize the acousticons in the old, accustomed way, regardless of which synthesizer or which set of sounds has the respective telephone service installed at the moment.

We introduce the classes of the acousticons and give their brief characteristics as well as several examples of sounds. The whole definition containing the full list of sound descriptions and names of the sound files exceeds the frame of this paper.

<table>
<thead>
<tr>
<th>Code</th>
<th>Class</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1-A16</td>
<td>Acoustic emoticons</td>
<td>sounds reflecting human feelings, moods and attitude to the text</td>
<td>short giggling, laughter, devil laughter, Oooops..., Wow!, Yeees!, sad groan … Sounds suitable for acoustic interpretation of the graphical emoticons.</td>
</tr>
<tr>
<td>B1 - B16</td>
<td>Babycons</td>
<td>Acoustic displays of children</td>
<td>children giggling, cry, etc. If You want to have your advertisement successful, use a child in it!</td>
</tr>
<tr>
<td>E1 - E16</td>
<td>Eroticons</td>
<td>Sounds of love, passion, sex, yearning</td>
<td>kisses, hard beating, sniff, screams, orgasm etc. A person in love does not worry about telephone fees.</td>
</tr>
<tr>
<td>V1 - V16</td>
<td>Vulgaricons</td>
<td>Indecent sounds, “dirty sounds” or sounds on the boundary of social acceptability</td>
<td>Fart, belch, spittle, vomit, squelch, hiccup, snore… Whether You like it or not, these sounds belong to the most marketable.</td>
</tr>
<tr>
<td>Z1 - Z16</td>
<td>Zooicons</td>
<td>Acoustic displays of animals</td>
<td>Roaster, dog, cat, horse, lion, hen, pig, goat, donkey, mouse, snake, gadfly… Which of the sounds would You send to your enemy?</td>
</tr>
<tr>
<td>I1 - I16</td>
<td>Symbolicons</td>
<td>Illustrative and symbolical sounds</td>
<td>Church bell, clocks, gun shot, circular saw, glass crack, doors, toilet, etc.</td>
</tr>
</tbody>
</table>
T1 - T16 Transporticons  Sounds of transport means and vehicles
Human steps, horse gallop, car alarm, car crash, car brakes, locomotive, firemen car, ambulance,
The sound of horse gallop can mean “I am hurrying to meet You!”

P1 - P16 Partycons  Sounds of party and having fun with friends
Mortise joint, pinging with glasses, opening a bottle of vine, opening a bottle of champagne, sipping, step dancing, Cheers, drunk singing

S1 - S16 Sportikons  Sports
Table tennis, tennis, judge’s whistle, gong, mountaineer falling from a rock, stadium atmosphere,

J1 - J16 Instrumenticons  Jingles or sounds played by musical instruments
Jaw harp, cymbal, church organ, drums, ethnic instruments

M1-M16 Melodicons  Fragments of the well known melodies with a symbolical meaning
Jingle bells, Happy birthday, Wedding march etc.

7.3. Code for inserting the sounds in the SMS messages.
The letter in the code designates the class and every sound in the class have its own number.
The second way of inserting the sounds is to remember the names of the soundfile, which is listed in the full definition of General SOUNDI specification (e.g. kiss1 = E1, or gallop = S2)

8. Conclusions
The practical use of the speech synthesizer in SMS to Voice telephone service has showed the need of implementation entertaining features in the speech synthesizer. We have designed several of them and we hope that at least some of these features will gain a popularity among the SMS to Voice users in the year 2004.

9. Acknowledgements
This research was supported by the Slovak Agency for Science VEGA, grant No.2/2087/22.

10. References

http://members.tripod.com/~ringtones/note_syntax.txt