THE CONSTRUCTION OF A MIXED UNIT INVENTORY FOR MACEDONIAN TEXT-TO-SPEECH SYNTHESIS

Branislav Gerazov and Zoran Ivanovski, Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia

Abstract - The paper presents in detail the construction process of the mixed inventory unit database used for the text-to-speech (TTS) system “Speak Macedonian”. Units called quasi-diphones form the basis of the system’s mixed inventory. The selection of the quasi-diphone units to be recorded from the complete diphone set of Macedonian is discussed in detail and a list of the chosen diphones is included in the paper. The selection was made to provide maximum synthesized speech quality at minimum inventory size. Details have been given of the recording and the unit segmentation process. The interface between the recorded audio and the TTS system was also discussed. The presented information may be of use for the development of future databases for TTS systems both for Macedonian and for other languages.

Key words – TTS, unit, mixed inventory, diphone, segmentation, recording

1. INTRODUCTION

The three paradigms of Text-to-Speech (TTS) synthesis are articulatory, formant and concatenative synthesis, [1 - 3]. Up until the 1990’s, TTS synthesizers predominately used articulatory and formant synthesis, both based on modeling the speech production process and thus both with a high development cost. Since then TTS systems became increasingly based on concatenative synthesis, which uses concatenation of prerecorded natural speech segments, also called units, to generate the requested speech output. This approach gives the synthetic speech a very natural sound while avoiding the expensive speech model development process, thus concatenative speech synthesis is extremely cost-effective and has allowed the proliferation of TTS systems across the globe.

Various unit lengths are used in concatenative synthesis, the longer the units the more natural the output speech, but the bigger the database. Most systems are based on diphones, which comprise two half-phones and the transition between them. Diphones usually number around 1000 per language and provide for reasonable speech quality. The most popular diphone based systems are AT&T’s diphone based TTS system and Dutoit’s MBROLA synthesizer [4, 5]. These systems are called monorepresented inventory systems because they mostly have only one recording of each unit to use for synthesis.

Modern state-of-the-art TTS system use expanded inventories with large unit databases with multiple units per phonetic content. The quality these systems relies heavily on the algorithms used for the selection of the most appropriate unit for concatenation, giving them the name unit-selection TTS systems. Famous unit-selection TTS systems are the Festival System from the University of Edinburgh and AT&T’s NextGen synthesizer [6, 7].

To date two high quality systems have been developed for TTS synthesis in Macedonian: TTS-MK, a diphone based system developed at FON University, [8], and “Speak Macedonian”, a mixed-length unit inventory system developed by the authors [9]. Both are monorepresented inventory systems.

The quality of the unit inventory is detrimental to the output speech quality of the TTS system, especially in monorepresented inventory systems. Because of this the creation of the inventory i.e. the recording and segmentation of the speech audio material is crucial to the overall quality of the TTS system, [10, 11, 12, 13].

This paper presents in detail the recording and segmentation process used in the creation of the unit inventory of the system “Speak Macedonian”. It is by disseminating this information that we hope to contribute to future unit inventory developments.

2. DIPHONES VS. QUASI-DIPHONES

Although the phone is the basic acoustic unit of speech, synthesizing speech by concatenating phones yields poor results. This is because of the great difficulty in simulating the interphone transition. The more common approach is to use a unit that includes half of each phone and the transition between them, called a diphone.

The TTS system “Speak Macedonian” uses a unique mixed-rank inventory featuring two sets of units: phones and quasi-diphones. The quasi-diphone (further written q-diphone) is a unit we created as a variation of the classic diphone concept. The q-diphone differs to the diphone in the respect that it comprises the two phones in their entirety. This allows q-diphones to be concatenated naturally with regular phones and vice versa.

When the need arises of concatenation of two succeeding q-diphones, q-diphones are easily demoted to classic diphones by trimming and can thus be concatenated. The trimming is facilitated using transition markers that give the boundary location between the two phones within each q-diphone, with the cut set to half way between the transition marker and the end to be trimmed.

The use of q-diphones decreased the necessary size of the inventory which in turn decreased the development cost of the inventory, while maintaining high-quality output.

3. MACEDONIAN PHONES

Macedonian is comprised of 34 basic phones, 28 of which are represented with a unique letter in the alphabet. The letters together with their IPA (International Phonetic Alphabet) equivalents are presented in Table I. Of these, five
are vowels (/a/, /b/, /i/, /o/ and /u/), and the rest are consonants. Grouped according to the manner of articulation the consonants are comprised of: 6 plosives (/b/, /p/, /g/, /k/,
/d/, /t/), 3 approximants (/j/, /r/), 7 fricatives (/v/, /f/, /z/, /s/,
/l/, /d/), 6 affricates (/d\, /n/, /m/), 3 nasals (/m/, /n/, /
/ng/) [14]. The letter “a” reads /æ/ but also /a/ when preceding the front vowels /ɛ/ and /i/, and the approximant /j/. The letter “n” reads /n/. The cluster /lj/ was treated as a phone, due to its compactness and the tendency to palatalize it in common speech. The phone /r/ can become syllabic /ɾ/ when enclosed by consonants, as it is in: “првиот” /prvi\ t/ (eng. the first), “пред” /p\v/ (eng. spine) etc. Finally /n/ is velarized to /ŋ/ before /k, g/, as in “банка” /ban\ka/ or “англиски” /angl\ski/, [15]. The phones /l/, /ɾ/, and /ŋ/ are coded with “w”, “q” and “n” further in the text.

### TABLE I

**PHONE INVENTORY OF MACEDONIAN WITH CORRESPONDING IPA EQUIVALENTS**

<table>
<thead>
<tr>
<th>A</th>
<th>/a/</th>
<th>Н</th>
<th>/i/</th>
<th>C</th>
<th>/s/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Б</td>
<td>/b/</td>
<td>Ј</td>
<td>/j/</td>
<td>Т</td>
<td>/t/</td>
</tr>
<tr>
<td>В</td>
<td>/v/</td>
<td>К</td>
<td>/k/</td>
<td>К</td>
<td>/c/</td>
</tr>
<tr>
<td>Г</td>
<td>/g/</td>
<td>Л</td>
<td>/l/</td>
<td>У</td>
<td>/u/</td>
</tr>
<tr>
<td>Д</td>
<td>/d/</td>
<td>Љ</td>
<td>/lj/</td>
<td>Ф</td>
<td>/f/</td>
</tr>
<tr>
<td>Е</td>
<td>/e/</td>
<td>М</td>
<td>/m/</td>
<td>Х</td>
<td>/x/</td>
</tr>
<tr>
<td>Ж</td>
<td>/z/</td>
<td>О</td>
<td>/o/</td>
<td>У</td>
<td>/u/</td>
</tr>
<tr>
<td>Ј</td>
<td>/y/</td>
<td>П</td>
<td>/p/</td>
<td>Ш</td>
<td>/ʃ/</td>
</tr>
<tr>
<td>Р</td>
<td>/r/, /ɾ/</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3. MACEDONIAN DIPHONES

The set of diphones found in a language is an essential starting point for developing a diphone-based TTS system for that particular language. This is highly specific to the language at hand, and calls for its thorough phonetic analysis. Such an analysis for the Macedonian language was undertaken by the authors as presented in [16]. In the analysis, a set of 707 unique diphones was extracted from a large body of text in Macedonian, that totaled more than 2.3 million words. This diphone set served as the basis for the recording of the q-diphone unit inventory used by our system.

### 3. UNIT DATABASE

#### A. Selecting the Quasi-Diphones

A selection of diphones was made from the complete set of diphones for the recording of the q-diphone set in the unit inventory. The following sets of diphones were eliminated from the final recording list:

i) diphones that began or ended with silence – 62 in total. This was directly provided by the use of q-diphones which allow using a simple fade-in/fade-out at the boundary of word instead of diphones comprising silence.

ii) diphones that ended with a voiceless plosive (/p/, /k/, /t/) or affricate (/t\, /ts/, /c/) – 114 in total. These phones are always preceded by a short period of silence due to their manner of articulation, i.e. there is a short stop in the airflow through the vocal tract which generates a build up of air pressure, that is released in a short burst to articulate the phone. This separates their waveform completely from the preceding phones, which have also little influence on their spectral content. This is because spectral coloring of a phone is largely influenced by the succeeding phone, because the vocal tract is already preparing to articulate it, termed coarticulation. The coarticulation effects of plosives on previous phones was seen as minor and disregarded. Fig. 1 shows spectral coloring for the plosive /k/ in two different phonetic contexts - /uka/ and /aku/. We can see that the coloring is influenced by the phones succeeding /k/ and not preceding. Also we can see that /k/ itself has little influence on the spectral content of its predecessors.

iii) diphones that ended with a voiced plosive (/b/, /g/, /d/), or affricate (/d\, /dz/, /ts/) – 87 in total. In contrast to voiceless plosives and affricates these phones do not feature a complete stop in vocalization. Rather during the pressure buildup process airflow is maintained through the nasal cavity giving a nasal sound. Because of the low level of this nasal sound a smooth transition to it was seen as unnecessary to maintaining a high output speech quality. Fig. 2 illustrates the spectral behavior of /g/ in two phonetic contexts: /uga/ and /agu/. We can see that although /g/ is heavily influenced by the following phoneme it has little influence on its predecessors.

iv) diphones that comprised two identical phones – 12 in total. These can be easily obtained using the phone set.

Other q-diphone groups where also considered for elimination such as those containing fricatives, however due to coarticulation effects, they were included in the database, Fig. 3. This increased the database from the original 131 to 449 q-diphones, giving a reduction of 36% from the 707 full diphone inventory size. The recorded q-diphones are given in Table II at the end of this paper.
B. Recording and segmentation

The set-up used for recording the audio needed for building the database featured a standard consumer microphone connected to a Lexicon Lambda audio interface, and a personal computer with audio editing software. The microphone was positioned at a 25 cm distance from the speaker at a 30° angle from the speaker’s central axis. This gave a good signal quality eliminating puffs from plosives and high levels from fricatives, while maintaining an acceptable signal level. The digitization was done with a 44.1 kHz sampling rate and a 16-bit resolution.

The total set of q-diphones was recorded in several sessions over a 3-day period. Each of the sessions was divided into recording time and segmentation. One of the authors’ voices was used for the recording process. The units were extracted from direct unit vocalization, not from continuous speech. This allowed for a much more efficient recording and inventory creation process. A short subset (up to 10) diphones were spoken and recorded at a time, each of them spoken a minimum of 2, but for some diphones up to 10 times. Special attention was put on the clear pronunciation of the diphones at hand while maintaining normal speaking rhythm. The latter was important because putting the speaking focus on the diphones increased their duration in respect to their duration in normal speech. This was seen as problematic during the segmentation of some phones, such as plosives and affricates (/k/, /tʃ/ etc.) which needed to be preserved in their entirety.

During pronunciation the q-diphones were enclosed in carrier syllables with no semantic meaning that were made up by the speaker during the recording process. These carrier syllables mainly comprised of two vowels added on each end of the q-diphone, except for the q-diphones that contained a vowel already. This was necessary in order to get proper amplitude levels for the diphones, which were not provided when the q-diphone was used at the beginning or the end of an utterances. Devoicing also precluded the use of some q-diphones at the end of utterances. Pitch frequency of the q-diphones was maintained by playing a reference 120 Hz drone to the speaker in the recording process.

These short recordings were then processed with a denoising algorithm based on spectral subtraction with a setting of -6 dB noise reduction at a +3 dB offset to the predicted noise level. This was necessary to eliminate the high frequency noise present in the recording due to the low quality microphone used.

After each diphone subset was recorded and denoised, the recording was analyzed and the best quality diphones were selected for segmentation. These were copied together with their surrounding phones to a master wav audio file where they were normalized. Normalization was based on the vowels as they are the loudest and were included in all carrier syllables. The following normalization levels were used: -8 dB for open and mid vowels (/a/, /ɛ/, /i/), -10 dB for closed vowels (/ɔ/, /u/), and -12 dB if the previous choices amplified the diphone of interest too much, e.g. when surrounding vowels were spoken at a lower level than usual. Some fricatives, such as /ʃ/, were normalized on their own to lower their level.

Finally, each q-diphone was marked with three markers. The first gives the start point of the first of the phones in the pair. This marker also holds the name of the q-diphone. The second marker gives the transition point between the phones, and the third the end of the q-diphone, Fig. 4. The marker data is saved in a separate mrk file, that is parsed by the system to extract the units from the audio file. The use of markers and a single audio file, allow easy database management. All unit corrections can now be done with simple marker repositioning, and the addition of new q-diphones is easily achieved by adding the q-diphone to the audio file and marking it. The TTS system automatically applies the changes to its inventory at run-time.
During the marker positioning special attention was paid to the unit length. As mentioned, due to the deliberate focus on the diphones, recorded phone lengths were greater than their average lengths in normal speech. Several unit lengths were used as standard for segmentation: 100 ms for vowels, 80ms for fricatives and affricates, 70ms for approximants, and 60ms for the rest. Some units were necessarily included whole such as plosives and affricates which were segmented with an included 30 ms stop duration. The absolute unit duration in the inventory is not detrimental to the rhythm of the output speech as unit lengths are modified using the pitch-synchronous overlap-add (PSOLA) algorithm during synthesis, however, the closer they are to the length required in synthesis the better the speech output quality.

At the end representative segments for each phone were chosen from the master audio file and were copied to another audio file that would comprise the phone set. The set contains 32 phones. The phones /v/ and /y/ were not included, as they appear only in the diphones already contained in the q-diphone set. Two markers were used to select the phones.

The master audio file has a total duration of 3 min 50 s with a total q-diphone unit duration of 1 min 20 s, at a 175 ms q-diphone duration average. The size of the wav file is 19.4 MB, of which 6.63 MB go to the q-diphons. This gives a 34% efficiency which is understandable giving that q-diphons are usually selected from parts of the phone durations and other phones are included in the file as well as periods of silence. The second wav file has a total duration of 8.5 s with 2.6 s of phones at 30% efficiency, averaging to 82 ms per phone. The file size is 742 KB, of which 227 KB go to the phones. This data is summarized in Table III.

The TTS system used the marker data to create the unit inventory from the audio file. The names of the segments are extracted from the start markers and stored in a name matrix. No list of q-diphones is required in the TTS system for the inventory to be built. This allows easy inventory scalability. The audio for each q-diphone is extracted from the start marker to the stop marker and stored in a cell array. The positions of the transition markers for the q-diphones are also stored for use in the synthesis algorithm. Namely if one phone is mapped to two consecutive q-diphones such as /ɔ/ and /n/ are in the q-diphone mapping of the word /mak6dɔnski/ i.e. /ma/-/k6/-/dɔ/-/ɔn/-/ns/-/ni/, then the q-diphones /dɔ/, /ɔ n/ and /ns/ are degraded to classic diphones at the borders where they meet by disregarding the halves of the phones at borders relative to the transition markers. Transition markers are also used in the PSOLA algorithm to determine which part of the q-diphone is voiced and which is unvoiced, which is important for the pitch and duration modification process.

### Table III – Q-Diphone and Phone Unit Set Data Summary

<table>
<thead>
<tr>
<th></th>
<th>unit set</th>
<th>q-diphone</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>wav total</td>
<td>3 min 50s</td>
<td>8.5s</td>
<td></td>
</tr>
<tr>
<td>filesize</td>
<td>19.4 MB</td>
<td>742 KB</td>
<td></td>
</tr>
<tr>
<td>wav unit</td>
<td>1 min 20s</td>
<td>2.6s</td>
<td></td>
</tr>
<tr>
<td>filesize</td>
<td>6.63 MB</td>
<td>227 KB</td>
<td></td>
</tr>
<tr>
<td>unit count</td>
<td>449</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>average duration</td>
<td>175 ms</td>
<td>82 ms</td>
<td></td>
</tr>
<tr>
<td>peak level [dB]</td>
<td>-5 dB</td>
<td>-7 dB</td>
<td></td>
</tr>
<tr>
<td>RMS power [dB]</td>
<td>-13 dB</td>
<td>-15 dB</td>
<td></td>
</tr>
<tr>
<td>minimum power [dB]</td>
<td>-35 dB</td>
<td>-52 dB</td>
<td></td>
</tr>
<tr>
<td>average power [dB]</td>
<td>-21 dB</td>
<td>-23 dB</td>
<td></td>
</tr>
</tbody>
</table>

### 4. CONCLUSION

The paper presents in detail the creation process of the mixed inventory unit database used for the TTS system “Speak Macedonian”. The unit inventory is based on q-diphone units a set of which was selected from the complete diphone set of Macedonian. The criteria used in the selection process are discussed in detail and a list of the recorded q-diphones is included in the paper. The selection made provides maximum speech quality at minimum inventory size. Details have been given of the recording and the unit segmentation process. At the end, the interface between the recorded audio and the TTS system was also discussed. The presented information may be of use for the development of future databases for TTS systems both for Macedonian and for other languages.

### REFERENCES


Божковска С., Л. Минова-Гурков, Д. Пандев и Ж. Цветковски, Оптика Ираматика на Македонскиот Јазик, Просветно Дело, 2008.
