“Hanmal” Korean Language Diphone Database for Speech Synthesis

Hyunsong Chung*

ABSTRACT

This paper introduces a “Hanmal” Korean language diphone database for speech synthesis, which has been publicly available since 1999 in the MBROLA web site and never been properly published in a journal. The diphone database is compatible with the MBROLA programme of high-quality multilingual speech synthesis systems. The usefulness of the diphone database is introduced in the paper. The paper also describes the phonetic and phonological structure of the database, showing the process of creating a text corpus. A machine-readable Korean SAMPA convention for the control data input to the MBROLA application is also suggested. Diphone concatenation and prosody manipulation are performed using the MBR-PSOLA algorithm. A set of segment duration models can be applied to the diphone synthesis of Korean.

Keywords: diphone database, MBROLA, speech synthesis, Korean, duration

1. Introduction

Dutoit et al. (1996) point out that the ability of concatenative synthesizers to produce high quality speech is dependent on the type of segments chosen and the model of speech signal to which the analysis and synthesis algorithms refer. The design should be able to account for as many co-articulatory effects as possible. Given the restricted smoothing capabilities of the concatenation technique, they should be easily connectable. Their number and length should also be kept as small as possible.

In order to create synthetic speech manipulated by a temporal model and to evaluate its perceptual quality, a new Korean language diphone database “Hanmal” was developed by Chung, Huckvale, and Kim (1999) based on the MBROLA synthesis system. This diphone database has been publicly available since September 17, 1999 from the MBROLA web site (Dutoit et al., 1996) so that other researchers could synthesize the Korean speech and investigate the relationships between prosody variation and naturalness. MBROLA is a speech synthesis system based on the concatenation of diphones. It takes a list of phones as input, together with prosodic information (duration of phones and a piecewise

* Department of English Education, Korea National University of Education

linear description of pitch), and produces speech signals, at the sampling frequency of the diphone database used.

The proper documentation of the diphone database has not been available and never been published in a journal. The aim of this paper is to provide the general public with the proper documentation of the database.

2. Creating a Text Corpus

Diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the following one. Their main usefulness in synthesis is that they minimize concatenation problems, since they contain most of the transitions and co-articulations between phones. They also require relatively small amounts of memory, as their number remains small (compared to synthesis units such as half-syllables or triphones).

The first step in building a diphone database is to generate a list of all the phones of the language. Notice that phones are acoustic instances of phonemes. To obtain a list of phones from a list of phonemes requires the investigation of which acoustic versions of phonemes differ significantly due to co-articulation. Although it is not necessary to account for all allophonic variations to build an intelligible synthesizer, the naturalness of synthetic speech may be affected if too few allophones are considered. When a complete list of phones has emerged, a corresponding list of diphones is readily obtained, and a list of words can be constructed such that each diphone appears at least once.

To prepare a diphone database capable of satisfying these requirements for Korean, 1,986 nonsense words were created to cover a catalogue of 1,986 diphones. In order to make the database acceptable to the general public, the MBROLA project team asked the authors to use the SAMPA (Speech Assessment Methods Phonetic Alphabet; Wells, 2004) transcription convention. Upon the request of the MBROLA team, a Korean SAMPA convention were developed. SAMPA is a machine-readable phonetic alphabet. SAMPA basically consists of a mapping of symbols of the International Phonetic Alphabet onto ASCII codes. The Korean SAMPA (K-SAMPA) has been further developed by Kim (2001). However, the database in this paper used the convention in its original format. Table 1 and 2 list the consonants used in the diphone database in IPA and SAMPA notation.

Table 1. List of onset segments.

IPA	SAMPA	DESCRIPTION
k	k	velar lax stop, voiceless
k'	k_>	tense velar stop
n	n	alveolar nasal
t	t	alveolar lax stop, voiceless
t'	t_>	tense alveolar stop
ɾ	4	alveolar tap
m	m	bilabial nasal
p	p	bilabial lax stop, voiceless
p'	p_>	tense bilabial stop
s	s	alveolar fricative
s'	s_>	tense alveolar fricative
ŋ		not assigned
ts	ts\	alveo-palatal lax affricate, voiceless
ts'	ts_>	alveo-palatal tense affricate
ts^h	ts_h	aspirated alveo-palatal affricate
k^h	k_h	aspirated velar stop, voiceless
t^h	t_h	aspirated alveolar stop
p^h	p_h	aspirated bilabial stop
h	h	glottal fricative
g	g	velar stop, voiced
d	d	alveolar stop, voiced
b	b	bilabial stop, voiced
dz	dz\	alveo-palatal affricate, voiced
ɕ	s\	alveo-palatal fricative
l	l	alveolar lateral

Table 2. List of coda segments.

IPA	SAMPA	DESCRIPTION
k̚	k_}	velar stop, voiceless, no audible release
n	n_}	alveolar nasal
t̚	t_}	alveolar stop, voiceless, no audible release
l	l_}	alveolar lateral
m	m_}	bilabial nasal
p̚	p_}	bilabial stop, voiceless, no audible release
ŋ	N	velar nasal

The consonants were grouped into 19 onset consonants and 7 coda consonants, because the developers believed that Korean listeners were likely to be sensitive to unreleased consonants occurring in coda position. The position of consonants in the syllable is determined based on the phonetic form of the utterance. So when the syllable final consonant is not resyllabified, then it is in the coda position; when it is resyllabified, it should be in the onset position. In order to distinguish coda consonants from syllable onset consonants, the unreleased stop diacritic "_}" was appended to coda consonants "k", "n", "t", "m", "l" and "p". Allophonic variants of consonants were then established as a function of their segmental context. For instance, every lax obstruent stop and affricate was matched with its voiced counterpart. The lax velar stop has two allophones in the onset position: voiceless "k" and voiced "g". If the segment follows a voiced segment, it becomes voiced. In the coda position, it becomes "k_}". The alveolar stop has "t" and "d" in the onset position, "t_}" in the coda position. The bilabial stop has "p", "b" and "p_}". The lax alveo-palatal affricate also has two allophones: "ts\" and "dz\" in the onset position, but in the coda position they are neutralized to "t_}". The lax alveolar fricative has two allophones in onset position: "s\" before a high vowel and "s" otherwise. Among obstruents, tense unaspirated stops, tense aspirated stops and fricatives are all neutralized in the coda position. Alveolar/alveo-palatal obstruents "ts_h", "ts_>", "t", "t_>", "s_>", and "s" are neutralized to "t_}"; velar obstruents "k" and "k_}" are neutralized to "k_}"; bilabial obstruents "p" and "p_>" are neutralized to "p_}"; the glottal fricative "h" is neutralized to "t_}". None of these obstruents in coda position have voiced equivalents. Among sonorants, "n", "l", and "m" appear in syllable initial position. "l" has an allophone "4" when it appears in intervocalic position. However, the phonemic status of Korean liquids are controversial. This database follows Kim-Renaud (1974) and Ahn (1985) where the lateral is more widely considered as the underlying phoneme than the flap. Though phonologically, "N" can appear in the syllable initial position, it rarely appears in that position. So "N" was put in the coda position. In the coda position, sonorants can be "n_}", "l_}", "m_}" and "N".

Table 3 lists the vowels used in the diphone database in IPA and SAMPA notation. Korean vowels consist of 9 monophthongs and 12 diphthongs in this convention. Each diphthong was treated as a unitary segment in the diphone database. Because there are no significant variations of vowel realization in context, no allophonic variants of vowels were considered.

From this list of segments, 12 groups of nonsense words were constructed to define all the available diphone contexts. Group 1 covers all the voiced syllable onset consonants in combination with following vowels. Group 2 covers all vowel to vowel combinations, while Group 3 all vowel and coda consonant combinations, and Group 4 all vowel and pause combinations. Other groups covered coda consonant and onset consonant combinations, vowel and onset consonant combinations, syllable coda consonant and pause combinations, pause and onset consonant combinations, pause and vowel combinations, voiceless onset consonant and vowel combinations, coda and vowel combinations, and pause alone. A list of the groups and their size is shown in Table 4.

Table 3. List of vowel segments.

IPA	SAMPA	DESCRIPTION
a	a	open front unrounded
ɛ	E	open-mid front unrounded
ja	ja	palatal approximant + open front unrounded
jɛ	jE	palatal approximant + open-mid front unrounded
ʌ	V	open-mid back unrounded
e	e	close-mid front unrounded
jʌ	jV	palatal approximant + open-mid back unrounded
je	je	palatal approximant + close-mid front unrounded
o	o	close-mid back rounded
wa	wa	voiced labial-velar approximant + open front unrounded
wɛ	wE	voiced labio-velar approximant + open-mid front unrounded
ø	2	close-mid front rounded
jo	jo	palatal approximant + close-mid back rounded
u	u	close back rounded
wʌ	wV	voiced labio-velar approximant + open-mid back unrounded
we	we	voiced labio-velar approximant + close-mid front unrounded
wi	wi	voiced labio-velar approximant + close front unrounded
ju	ju	palatal approximant + close front rounded
ɯ	M	close back unrounded
ɰi	M\i	velar approximant + close front unrounded
i	i	close front unrounded

Table 4. Diphone groups in contexts.

GROUP	COMBINATION	NUMBER
1	(voiced) onset + nucleus	378
2	nucleus + nucleus	441
3	nucleus + coda	147
4	nucleus + pause	21
5	coda + onset	133
6	nucleus + onset	399
7	coda + pause	7
8	pause + onset	18
9	pause + nucleus	21
10	(voiceless) onset + nucleus	399
11	coda + nucleus	21
12	pause + pause	1
Total number of diphone used		1,986

3. Recording

The speaker was 48 years old at the time of the recording and a native Korean male who was born in Busan, Korea. He had studied and lived in Seoul for 7 years and in America for 10 years. The recordings were made four times in an anechoic chamber on digital tape using 2 channels at 44,100 samples/sec/channel. Channel 1 was the speech signal from microphone, channel 2 was a Laryngograph signal. They were resampled to 16 kHz and transferred to disk. In order for the MBROLA resynthesis operation to achieve best results, the corpus was read with a monotonous intonation. The speaker was also requested to keep the pitch and rhythm consistent across phrases. This consistency aids in the production of smooth segment concatenation. However, in order to improve the naturalness of speech made from the diphone database, the speaker was requested to read each nonsense phrase rapidly and fluently. In order to avoid any vocal fry in the diphone database, a neutral vowel /ə/ was inserted before the target words except for those starting with a pause or a voiceless consonant.

4. Annotation

The Speech Filing System (SFS; Huckvale, 2004) was used to analyze and annotate the speech data. The segmentation was decided with reference to three signals: waveform, spectrogram, and Laryngograph signal (Lx). Three boundary points were identified: the mid-point of each target segment and the boundary between the two target segments. Annotations were stored as sample numbers in a database and then exported in a text file for diphone processing. They look like the following:

a. a-a.d16	a	a	4526	7374	5844
b. a-ae.d16	a	E	5148	7757	6306
c. a-b.d16	a	b	3741	5334	4868
d. a-bb.d16	a	p'	2874	4971	3619
e. a-bc.d16	a	p_}	4274	6918	5346
f. a-ch.d16	a	ts_h	2342	4443	3062

As mentioned in section 2, diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the following one. In the above string, *.d16 refers to the speech signal data filename. Segments in the second and third columns are the target diphones. The fourth column is the starting sample number of the diphone and the next column is the end point of the diphone. The last column indicates the mid point of the diphone, that is, the boundary between two target segments. The boundary of two target segments were manually annotated. The mid point does not necessarily correspond to the very half of the whole diphone sample numbers.

5. MBROLA Program

The diphone recordings were processed by the MBROLA team in Belgium to produce the “Hanmal” diphone database. Applications based on this database are supported on a wide range of computing platforms using the MBROLA signal generation engine. Diphone concatenation and prosody manipulation can be performed using the MBR-PSOLA algorithm (Dutoit et al., 1996). This method is an interesting alternative to purely time-domain PSOLA, in the context of a multi-lingual TTS system, for which the ability to derive segment databases automatically, to store them in a compact way, and to synthesize high quality speech with a minimum number of operations per sample is of considerable interest. The format of the control data input to the MBROLA application is as follows. The target word is “kan_}da (to go)”.

```
_    100
k    35
a    79 20 140 50 135 80 135
n_}  120
d    70
a    150 20 135 50 140 80 135
_    100
```

In the above string, “_” stands for the pause. The second column of each row represents the duration of the target segment in milliseconds. The other columns describe the pitch contour for the segment in pairs of numbers: the first value in the pair is the percentage position through the segment, the second value is the fundamental frequency in hertz. Pitch values are linearly interpolated inside and across segments. The input transcription needs to be fully specified for allophonic variants. For example, for the input /halapʌtsi/ (grandfather)" the file contains “_ h a 4 a b V dz\ i _" not "_ h a l a p V ts\ i _”. Letter-to-sound conversions can be carried out by using Jang's (2000) “Romanize” and “pronounce” scripts. However, the automatic conversion has not still been implemented in the application from the MBROLA web site due to technical problems.

6. Application to Speech Synthesis

A pronunciation dictionary is necessary to convert orthographic characters into the symbols used in this diphone database. Using a set of phonological rules, a lexicon has been constructed, which contains actual pronunciations of words. Each pronunciation is encoded in the lexicon as a metrical structure comprising syllable, onset, rhyme, nucleus and coda nodes as well as the segments, which are described using features. Phrases can be constructed from such a lexicon by concatenation of the prosodic structures and these may then be processed by rules of phonetic interpretation. This framework for prosodic synthesis follows that established by the ProSynth project (Hawkins et al. 1998). From the interpreted structure, a mapping can be made from the predicted phonetic properties, timing and intonational features to actual values input to the MBROLA application. Further results of application to speech synthesis and the perceptual evaluation of the synthesized speech manipulated by this diphone database can be found in Chung (2002) and Chung (2003).

7. Conclusion

This paper introduced a Korean language diphone database, "Hanmal", for speech synthesis. It showed the process of creating a text corpus for constructing the diphone database. A catalogue of 1,986 diphones based on 12 groups of nonsense words were created and a Korean SAMPA convention was also developed. The diphones were recorded by one Korean male speaker with monotonous but fluent speed to produce smooth segment concatenation. The speech and laryngograph signal were extracted from the recordings. The database has been publicly available since 1999 in the MBROLA web site. The diphone database is compatible with the MBROLA programme of high-quality multilingual speech synthesis systems. The paper described the phonetic and phonological structure of the database and how it was recorded and processed. A set of segment duration models were applied to the diphone synthesis of Korean.

References

Ahn, S-C. 1985. *The Interplay of Phonology and Morphology in Korean*, Doctoral dissertation, University of Illinois.

Chung, H. 2002. "Perceptual evaluation of duration models in spoken Korean." *Korean*

Journal of Speech Sciences, 9(1), 207-215.

Chung, H. 2003. "Sums-of-Products models for Korean segment duration prediction." *Korean Journal of Speech Sciences,* 10(4), 7-21.

Chung, H., Huckvale, M. & Kim, K. 1999. "A new Korean speech synthesis system and temporal model." *Proceedings of 16th International Conference on Speech Processing*, 1, 203-208.

Dutoit, Thierry, Pagel, V., Pierret, N., Bataille, F., and van der Vreken, O. 1996. "The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes." *Proceedings of 4th ICSLP,* Philadelphia, 3, 1393-1396. http://tcts.fpms.ac.be/synthesis/mbrola.html

Hawkins, S., House, J., Huckvale, M., Local, J., and Ogden, R. 1998. "ProSynth: an integrated prosodic approach to device-independent natural-sounding speech synthesis." *Proceedings of 5th ICSLP*, Sydney, 1707-1710.

Huckvale, M. 2004. http://www.phon.ucl.ac.uk/resource/sfs/

Jang, Tae-yeoub. 2000. *Phonetics of Segmental F_0 and Machine Recognition of Korean Speech*, Ph.D. thesis, University of Edinburgh.

Kim, Jong-mi. 2001. "Computer codes for Korean sounds: K-SAMPA." *Journal of Acoustical Society of Korea*, 20, 4E, 1-14.

Kim-Renaud, Young-Key. 1974. *Korean Consonantal Phonology*, Doctoral dissertation, University of Hawaii.

Wells, J. 2004. http://www.phon.ucl.ac.uk/home/sampa/

received: January 30, 2005
accepted: February 10, 2005

▲ Hyunsong Chung
Department of English Education
Korea National University of Education
San 7, Darakri, Gangnaemyeon, Cheongwongun
Chungbuk 363-791, Korea
Tel: +82-43-230-3554 (O)
E-mail: hchung@knue.ac.kr