A New Korean Speech Synthesis System and Temporal Model

Hyunsong Chung, Mark Huckvale*, Kyongsok Kim***
* Department of Phonetics and Linguistics, University College London
Gower Street, London WC1E 6BT, United Kingdom
** Department of Computer Science, Busan National University
Busan 609-735, South Korea
e-mail: * {hchung, mark}@phon.ucl.ac.uk, ** gimgs@asadal.cs.pusan.ac.kr
Phone: *+44 207 504 5026, +44 207 504 5002, **+82 51 510 2292

ABSTRACT

This paper introduces a new publicly-available Korean diphone database for speech synthesis and reports on our latest work towards a model of Korean prosody. The diphone database is compatible with the MBROLA programme of high-quality multilingual speech synthesis systems. The first part of the paper describes the phonetic and phonological structure of the database and describes how it was recorded and processed. The second part of the paper reports on progress towards a model of segmental timing compatible with diphone synthesis of Korean. So far we have built a model of vowel duration based on the analysis of over 1000 syllables annotated for their segmental and supra-segmental contexts. Through the use of an automated search and error minimisation procedure we have estimated the parameters of a nine-factor model which explains over 80% of the variance in vowel duration in the training data.

1. INTRODUCTION

Contemporary speech synthesis systems provide good segmental quality through the use of concatenative signal generation methods. Such methods shift the main focus of research onto the prediction of intonation and timing from text. However, while English, Japanese and the main European languages can call on extensive previous research into prosody, there have been few studies of Korean prosody relevant for synthesis. Research work in this area would be stimulated by access to a high-quality speech signal generation system for Korean which could be driven from phonological and phonetic parameters. Using such a system, perceptual studies could be undertaken of the acceptability of temporal and intonational models.

There are a number of commercial Korean speech synthesis systems, among them, Hansori from Korea Telecom (KT), Keulsori from Korean Electronics and Telecommunications Research Institute (ETRI), MagicVoice from Samsung, and one from LG. However, because the speech databases used in these systems are not available to the public, they are not suitable as a basis for experimental investigations into Korean prosody. To remedy this situation we have developed a new Korean diphone database based on the MBROLA system [1]. We will make this database publicly available free of restrictions on use in the near future. The database, recorded from a single male native speaker of Korean consists of 1,692 diphones. Preliminary evaluations have been made by comparing its output (with natural prosody imposed) against fully synthetic speech from KT and ETRI. Most listeners are satisfied with the segmental quality of our system: nearly half thought it superior.

In section 2 of this paper we describe how the database was constructed, while in section 3 we describe our first attempts at the construction of a model of Korean prosody which operates automatically from a phonological representation of a phrase.

2. KOREAN DIPHONE DATABASE

2.1 Diphone Database

Dutoit *et al* [2] 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.

To prepare a diphone database capable of satisfying these requirements, we designed a catalogue of 1,692 diphones. In Korean, there are 19 consonant phonemes and 21 vowel phonemes which are clearly reflected in Korean alphabets. In order to make the database acceptable to the general public, we followed the system for transliteration of Korean script into Latin characters agreed in 1997 between South and North Korean delegates (ISO TR 11941). In order to distinguish the non-ambisyllabic syllable final consonants from syllable initial consonants we appended the diacritic symbol “c” to coda consonants “g”, “n”, “d”, “m”, “l” and “b”. We also used the diacritic symbol “v” to indicate voicing after the consonants “g”, “d”, “b” and “j”. However, end-users do not have to input these diacritics since we have also provided software to make the appropriate substitutions automatically.

We grouped the consonants into 19 in syllable initial (onset) position and 7 consonants in syllable final (coda) position. In our database, a coda consonant is a non-ambisyllabic consonant occupying a syllable final position in a closed syllable. When the consonant is ambisyllabic with the following syllable and it occupies the onset position of the following syllable, we treat it as an onset consonant. Allophonic variants of consonants were then established as a function of their segmental and supra-segmental context. For instance, every lax unaspirated obstruent stop and affricate has its voiced equivalents. Where there is a contrast between voiced and voiceless obstruents, the basic (underlying) segment is a voiceless one. The lax unaspirated velar stop has two allophones in the onset position: voiceless “g” and voiced “gv”. If the segment follows a voiced segment, it becomes voiced. In the coda position, it becomes “gc”. The alveolar stop has “d” and “dv” in the onset position, “dc” in the coda position. Bilabial has “b”, “bv” and “bc”. The lax unaspirated alveopalatal affricate also has two allophones: “j” and “jv” in the onset position. In the coda position, “j” is neutralized to “dc”. The lax fricative has two allophones in onset position: “sh” before a high vowel and “s” otherwise. Among obstruents, tense unaspirated and tense aspirated stops, and fricatives are all neutralized in the coda position. Alveolar/palatal obstruents “ch”, “jj”, “t”, “dd”, “ss”, and “s” are neutralized to “dc”; velar obstruents “k” and “gg” are neutralized to “gc”; bilabial obstruents “p” and “bb” are neutralized to “bc”; pharyngeal fricative “h” is neutralized to “dc”. None of these obstruents have voiced equivalents. Among the sonorants, “n”, “r”, and “m” appear in syllable initial position. “r” has an allophone “l” when it follows the “l” coda. Though “ng” can phonologically appear in the syllable initial position, it is rarely likely to appear in that position. So we put “ng” in the coda position. In the coda position, sonorants have “nc”, “lc”, “mc” and “ng”.

Korean vowels consist of 9 monophthongs and 12 diphthongs. Each diphthong is treated as a unitary segment in the diphone database, without splitting it into two vowels. Because there are no significant variations of vowel realisation in context, we did not consider any further allophonic variants for vowels. Table 1 lists the consonants and vowels used in the diphone database.

From this list of segments, 12 groups of nonsense words were constructed to define all the available diphone contexts. Group 1 consists of all the voiced syllable onset consonants in combination with following vowels. Group 2 consists of all vowel to vowel combinations, Group 3 all vowel and coda consonant combinations, Group 4 all vowel and pause combinations. Other groups consisted of 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 groups and counts are shown in Table 2.

2.2 Recording

The speaker was a standard Korean speaker who had lived in Seoul for 32 years before coming to the UK to study in 1997. 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 to retain the segmental naturalness of the diphone database, the speaker was requested to read each nonsense phrase rapidly and fluently to simulate a real utterance. The speaker was also requested to keep the pitch and rhythm consistent. This consistency is a prerequisite in the production of smooth segmental concatenation. In order to avoid any vocal fry in the diphone database, we put a neutral vowel “eo” before the target words except for those starting with a pause or a voiceless consonant.

2.3 Annotation

We used the Speech Filing System (SFS) [3] to analyze and annotate the speech data. The segmentation was decided with reference to three signals: waveform, spectrogram, and Laryngograph signal (Lx). We identified three boundary points: 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.d16     a    a     4526    7374    5844
a-ae.d16    a    ae    5148    7757    6306
a-b.d16     a    b     3741    5334    4868
a-bb.d16    a    bb    2874    4971    3619
a-bc.d16    a    bc    4274    6918    5346
```

```
a-ch.d16   a   ch   2342   4443   3062
```

*.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 point 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.

2.4 MBROLA Application

The diphone recordings were processed by the MBROLA team in Belgium to produce the kr1 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 is performed using the MBR-PSOLA algorithm [2]. 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 “ganda (to go)”.

```
_  100
g   35
a   79 0 140 50 135 100 135
nc 120
d   70
a  150 0 135 50 140 100 135
_  100
```

In the above table, “_” 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. At this stage, the input transcription needs to be fully specified for allophonic variants. For example, when you input “halabeoji (grandfather)” into the file, you should type “_ h a r a b eo j i _” not “_ h a l a b eo j i _”. To overcome this problem, we have been developing a lexicon which contains the pronunciation of words, which is described in the next section.

2.5 Tools

As mentioned above, a pronunciation dictionary is necessary to convert orthographic characters into the symbols used in this diphone database. Using a set of phonological rules, we have constructed a lexicon 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. An example entry is given in Table 3. 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 [4]. 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.

2.6 Evaluation

Since a comprehensive temporal model is not yet available, evaluation of the diphone system has been limited. By using the mbrolign program [5], we have been able to copy the prosody of natural speech onto concatenated diphone strings. Comparisons between such synthetic utterances and equivalent, but fully synthetic, utterances from KT and ETRI seem satisfactory. For the comparison and evaluation, we chose two sentences. The first sentence was "Baramgwa haesnim'i seoro him'i deo sedago datugo iss'seubnida."; the other was "Urineun minjogjungheung'eui yeogsajeog sa'myeong'eul ddigo i'ddang'e tae'eonassda." We played the natural speech first, and randomly played three other synthetic speech from KT, ETRI, and ours to 10 subjects. They were fluent Korean speakers who are studying in London. The result showed that nearly 50 % of the subjects considered our synthetic speech was more intelligible than the other two synthesized speech. We concluded that the segmental quality of our diphone database is satisfactory. However, we also considered that the intelligibility of our synthesized speech partly owes to the copy of the natural prosody. After completing the temporal model we will evaluate our database without the use of natural prosody.

3. TEMPORAL MODEL

3.1 Training Corpus

In order to investigate what factors determine the variation in vowel duration, we recorded and analysed a corpus of read speech. For this study, 600 artificial utterances were designed and recorded by a single speaker. The utterances systematically explored both syllable position and syllable composition within a sentence frame containing nonsense monosyllable pairs. For example: /ikʌsɯn V | V soɾita/ was used to investigate inherent vowel duration; /ikʌsi CV(C) | CV(C) soɾita/ for consonantal influences on vowel duration; /ikʌsi CV | CVCVCVCV/ for prosodic influences on vowel duration. The recordings were made three 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 to make the speaker keep a consistent rate of speech, we used a prompting tool when recording. Sentences were displayed on a monitor screen at five second intervals so that the speaker could read each sentence with a regular rhythm. A total of 1,054 syllables were annotated. From these a table of vowel timing data was extracted comprising the duration and a description of the segmental and supra-segmental context in which each vowel was found. The context was encoded as a set of 27 factors, each of which could be said to be active or not for the vowel in question. The list of factors is given in Table 5.

3.2 Parameter Estimation of the Timing Model

The vowel durations and vowel contexts established from the training corpus were used to estimate the parameters of a simple multiplicative timing model. The model estimates the duration of a vowel as a function of the identity of the vowel (v) and the context (c) in which it is found:

$$d(v,c) = d_{min}(v) + [d_{inh}(v) - d_{min}(v)]*F(c)$$

where $d_{min}(v)$ is the minimum duration of the vowel v; $d_{inh}(v)$ is the inherent duration of vowel v - i.e. the duration found in a 'neutral' context; and F(c) is a compression factor based on the context independent of the vowel:

$$F(c) = \Pi f_i$$

where each compression factor f_i has a value that depends on one component of the context, for example f_0 represents the 'phrase-final syllable' context, which takes a value different to one in phrase-final contexts and a value equal to one elsewhere.

Although it is possible to hypothesise which contexts might influence vowel durations it is necessary to use an automated procedure to establish the relative importance of the compression factors and the best value for each factor. To establish the best model an automated procedure was constructed as described below. The procedure determined the best factors and the optimal factor values by minimising the squared error of prediction on the training data.

The process used the 1,054 vowel duration measurements labelled according to the 27 different binary contexts hypothesised as being relevant for vowel duration. Minimum and inherent durations were estimated from the distribution of durations for each vowel type, these are listed in Table 4. For each hypothesised context in turn the best model comprising a single factor was found using a function minimisation procedure [6]. This process identified the most significant context and the optimal factor value for a model of a single factor. The context causing the greatest reduction in squared error was then accepted and the search repeated for the best two factor model by testing each of the remaining 26 contexts in turn. The best second factor is then chosen and the process repeated for a third factor and so on until the squared error fails to fall by a significant amount, in this case at about nine factors. The result of this procedure is shown in Table 6. The final model of 9 factors explains over 80% of the variance in the training data.

From this result, we can produce a simpler equation to predict the vowel durations in the training data. We can simplify the chosen 9 contexts under 5 phonological categories as follows:

$$F(c) = PP * CM * AS * VOC * AMB,$$

where:

PP (Phrasal Position Factor) =
1.72, if the vowel is in the phrase-final position (f_0),
0.93, if the vowel is in the phrase-initial position (f_1),
1, elsewhere.

CM (Consonant Manner Factor) =
0.31, if the vowel is before a stop consonant (f_{12}),
0.26, if the vowel is before a nasal consonant (f_{14}),
0.33, if the vowel is before a fricative consonant (f_{13}),
0.73, if the vowel is before a liquid consonant (f_{15}),
1, elsewhere.

ASP (Aspiration Factor) =
0.82, if the vowel is after a strong aspiration consonant (f_4),
1, elsewhere.

VOC (Voicing Factor) =
0.33, if the vowel is after a voiced consonant (f_{25}),
1, elsewhere.

AMB (Ambisyllabicity Factor) =
1.59, if the vowel is before an ambisyllabic consonant (f_{17}),
1, elsewhere.

3.3 Comparison

Some comparisons between the actual vowel durations and the predicted durations according to this formula are shown in Table 7. The fit with the training data is, as might be expected, quite good.

4. CONCLUSION

This paper has introduced a new Korean diphone database and a temporal model of vowel duration in Korean. This diphone database kr1 is undergoing final adjustments and will be made available to the public later this year. The temporal model is based on a set of minimum and inherent durations for Korean vowels in combination with a set of phonological contexts. Together these components provide an environment which can foster further research into spoken Korean.

Future work will address the prediction of consonantal durations, the prediction of segmental quality changes in context, and the generation of intonation contours from marked text.

onset consonants		vowel		coda consonants	
Latin	allophones	Latin	allophones	Latin	allophones
g	g, gv	a	a	g	gc
gg	gg	ae	ae	n	nc
n	n	ya	ya	d	dc
d	d, dv	yae	yae	l	lc
dd	dd	eo	eo	m	mc
r	r, l	e	e	b	bc
m	m	yeo	yeo	ng	ng
b	b, bv	ye	ye		
s	s, sh	o	o		
ss	ss	wa	wa		
ngo	null	wae	wae		
j	j, jv	oe	oe		
jj	jj	yo	yo		
ch	ch	u	u		
k	k	weo	weo		
t	t	we	we		
p	p	wi	wi		
h	h	yu	yu		
		eu	eu		
		eui	eui		
		i	i		

Table 1. Segment index used in the diphone database.

Group	Combination	Number
Group 1	onset * nucleus	378
Group 2	nucleus * nucleus	441
Group 3	nucleus * coda	147
Group 4	nucleus * pause	21
Group 5	coda * onset	133
Group 6	nucleus * onset	399
Group 7	coda * pause	7
Group 8	pause * onset	18
Group 9	pause * nucleus	21
Group 10	onset * nucleus	105
Group 11	coda * nucleus	21
Group 12	pause * pause	1
Total diphone numbers		1692

Table 2. Diphone groups in contexts.
Onset in Group 1 is “the voiced onset * nucleus” combination.
Onset in Group 10 is “the voiceless onset * nucleus” combination.

```
<LEXICON>
<ENTRY ID="PARAM"><HW>param</HW>
<PRONSEQ>
<PRON ID="1"><IPA
ID="1">'paramc</IPA><SYLSEQ>
<SYL STRENGTH="STRONG" WEIGHT="LIGHT">
<ONSET STRENGTH="WEAK">
<CNS AMBI="N" CNSANT="N" CNSCOR="N"
CNSDOR="N" CNSLAB="Y" CONSTR="N" CONT="N"
NAS="N" SON="N" SPR="N" VOCCOR="N"
VOCDOR="N" VOCLAB="N">p</CNS>
</ONSET>
<RHYME CHECKED="N" STRENGTH="WEAK" VOI="Y"
WEIGHT="LIGHT">
<NUC CHECKED="N" LONG="N" STRENGTH="WEAK"
VOI="Y" WEIGHT="LIGHT">
<VOC COR="N" DOR="N" LAB="N"
OPN="Y">a</VOC>
<VOC COR="N" DOR="N" LAB="N"
OPN="Y">a</VOC>
</NUC>
<CODA VOI="N">
<CNS AMBI="Y" CNSANT="Y" CNSCOR="Y"
CNSDOR="N" CNSLAB="N" CONSTR="N" CONT="Y"
NAS="N" SON="Y" SPR="N" VOCCOR="N"
VOCDOR="N" VOCLAB="N" VOI="Y">r</CNS>
</CODA>
</RHYME>
</SYL>
<SYL STRENGTH="WEAK" WEIGHT="LIGHT">
<ONSET STRENGTH="WEAK">
<CNS AMBI="Y" CNSANT="Y" CNSCOR="Y"
CNSDOR="N" CNSLAB="N" CONSTR="N" CONT="Y"
NAS="N" SON="Y" SPR="N" VOCCOR="N"
VOCDOR="N" VOCLAB="N" VOI="Y">r</CNS>
</ONSET>
<RHYME CHECKED="N" STRENGTH="WEAK" VOI="Y"
WEIGHT="LIGHT">
<NUC CHECKED="N" LONG="N" STRENGTH="WEAK"
VOI="Y" WEIGHT="LIGHT">
<VOC COR="N" DOR="N" LAB="N"
OPN="Y">a</VOC>
<VOC COR="N" DOR="N" LAB="N"
OPN="Y">a</VOC>
</NUC>
<CODA VOI="N">
<CNS AMBI="N" CNSANT="N" CNSCOR="N"
CNSDOR="N" CNSLAB="Y" CONSTR="N" CONT="N"
NAS="Y" SON="Y" SPR="N" VOCCOR="N"
VOCDOR="N" VOCLAB="N" VOI="Y">mc</CNS>
</CODA>
```

```
</RHYME>
</SYL>
</SYLSEQ></PRON>
 </PRONSEQ>
</ENTRY>
</LEXICON>
```

Table 3. The structure of the lexicon.

Segment	$d_{min}(v)$	$d_{inh}(v)$	Segment	$d_{min}(v)$	$d_{inh}(v)$
a	82	154	oe	118	216
ae	67	170	yo	122	224
ya	89	190	u	37	166
yae	135	246	weo	139	238
eo	79	168	we	118	230
e	71	179	wi	90	190
yeo	144	240	yu	69	180
ye	144	250	eu	68	161
o	51	175	eui	83	175
wa	138	232	i	48	164
wae	133	236			

Table 4. Minimum and inherent duration of vowels.
$d_{min}(v)$ = minimum duration of the vowel
$d_{inh}(v)$ = inherent duration of the vowel

factor	factor description
f0	phrase-final
f1	phrase-initial
f2	phrase-second
f3	phrase-third
f4	vowel after strong aspiration consonant
f5	vowel after slight aspiration consonant
f6	vowel after no aspiration consonant
f7	vowel after fricative consonant
f8	vowel after stop consonant
f9	vowel after nasal consonant
f10	vowel after affricate consonant
f11	vowel after liquid consonant
f12	vowel before stop consonant
f13	vowel before fricative consonant
f14	vowel before nasal consonant
f15	vowel before liquid consonant
f16	vowel after ambisyllabic consonant
f17	vowel before ambisyllabic consonant
f18	vowel after bilabial consonant
f19	vowel after alveolar conosonant
f20	vowel after velar consonant
f21	vowel after alveopalatal consonant
f22	vowel before bilabial consonant
f23	vowel before alveolar consonant
f24	vowel before velar consonant
f25	vowel after voiced segment
f26	vowel before voiced segment

Table 5. Factors used in the training corpus.

Number of Factors	Add Factor	Squared Error	Variance %
0		5,902,000	100
1	f0	4,382,000	74.4
2	f12	2,716,000	46.1
3	f1	1,869,000	31.7
4	f14	1,386,000	23.5
5	f4	1,271,000	21.6
6	f25	1,234,000	20.9
7	f13	1,202,000	20.4
8	f17	1,156,000	19.6
9	f15	1,129,000	19.2

Table 6. Factor distribution.

	$d_{min}(v)$	$d_{inh}(v)$	PP	CM	ASP	VOC	AMB	d(v,c)	d(v)
kk a _	82	154	0	na	na	na	na	206	218
_ d e bc	71	179	2	12	na	na	na	102	100
_ b i s	48	164	2	13	na	na	17	105	96
b o _	51	175	0	na	na	na	na	264	269
p u	37	166	0	na	4	na	na	219	219

Table 7. Camparison between estimated vowel duration and actual vowel duration.
$d_{min}(v)$ = minimum duration
$d_{inh}(v)$ = inherent duration
d(v,c) = estimated duration
d(v) = actual duration
na = not applicable
Numbers in columns from PP to AMB are the context values (f_0, f_1, …).

REFERENCES

[1] Dutoit, T., V. Pagel, N. Pierret, F. Bataille, and O. van der Vreken, (1996), "The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes, " *Proc. 4th ICSLP '96*, Philadelphia, vol.3, 1393-1396.

[2] Dutoit, Thierry, Henri Leich, (1994), A comparison of four candidate algorithms in the context of high quality text-to-speech synthesis. *Proceedings of ICASSP '94*.

[3] http://www.phon.ucl.ac.uk/resource/sfs.html

[4] Hawkins, S., J. House, M. Huckvale, J. Local & R. Ogden, (1998), "ProSynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis," *Proc. 5th ICSLP '98*, Sydney, 1707-1710.

[5] http://tcts.fpms.ac.be/synthesis/mbrolign/

[6] Nelder, J.A. & R. Mead (1965) "A simplex method for function minimization," *The Computer Journal*, vol.7, The British Computer Society, 308-313.