Design and evaluation of prosodically-sensitive concatenative units for a Korean TTS system

Kyuchul Yoon

Division of English
Kyungnam University
449 Wolyong-dong, Masan, Kyungnam, South Korea

Abstract

This paper describes the design and evaluation of prosodically-sensitive concatenative units for a Korean text-to-speech (TTS) synthesis system. The diphones used are prosodically conditioned in the sense that a single conventional diphone is stored as different versions taken directly from the different prosodic domains of the prosodically labeled, read sentences. The four levels of the Korean prosodic hierarchy were observed in the diphone selection process, thereby selecting four different versions of each diphone: three edge diphones from the prosodic domains of the intonational phrase (IP), accentual phrase (AP) and prosodic word (PW), and a non-edge diphone from the domain of the prosodic word. Due to the size of the corpus that we employed, our system covers only 36.4% of the 6,503 possible diphones. A listening experiment designed to evaluate the quality of the diphone database showed that listeners preferred stimuli composed of prosodically appropriate diphones. We interpret this as supporting the view that segments carry prosodic domain information.

Key words: Korean, prosody, diphone, text-to-speech synthesis

1 Introduction

It is well known that the pronunciations of parts of words, i.e., speech segments, depend on their neighboring segments; this fact is incorporated into

---

1 This work was supported by Kyungnam University Foundation Grant 2007. The author would like to thank Professors Mary Beckman and Chris Brew at The Ohio State University and Kirk Baker, Eunjong Kong and eighty native speakers of Korean for their help on this project.
almost every text-to-speech and automatic speech recognition system built today. However, the fact that segmental pronunciation varies as a function of prosodic context, i.e. depending on where the segment is located with respect to the prosodic organization, of the language, is usually not incorporated in language technology systems. The goal of this work is to contribute to the growing literature on the relationship between segmental pronunciation and prosody, and evaluate whether implementing prosody-based strategies within speech synthesis is possible, and whether it provides improved comprehension of the synthesized speech.

To illustrate allophonic variation, we use an English example. In English, the voiceless plosives /p, t, k/ are unaspirated when they occur in a cluster with /s/, as in spat, stab and ski, but at the beginning of a stressed syllable not preceded by /s/, they are pronounced with aspiration, as in pat, tab and key. However, this is an oversimplification. Although the strongest degree of aspiration can be heard at the beginning of a syllable bearing the primary stress, a certain degree of aspiration can also be found in other contexts, e.g. at the beginning of the unstressed syllable in career or in the word-final cluster with /s/ in task. Although these three variants of aspirated /k/ differ slightly in their acoustic representation, they are all aspirated relative to the /k/ in ski. If the degree of aspiration were measured in terms of the voice onset time (VOT), its distribution would not be fixed but variable. That is, the distribution would lie on a continuum of VOT values — a continuum of degrees of aspiration much like the continuum of degrees of palatalization triggered by the different following vowel contexts in key, cow, car, and coat.

Allophonic processes also apply across word boundaries. For example, English fricatives can become devoiced in certain contexts. The fricatives /v, z/ in love to go and has to go can become devoiced under the influence of the following voiceless obstruent /t/. While this process of devoicing before /t/ might be understood as a kind of coarticulation, not all examples of devoicing of fricatives can be modeled in this way. The sibilant fricative /z/, especially, can be produced with varying degrees of devoicing even when there is no following voiceless segment, and the amount of devoicing changes depending on where the segment is located with respect to the syllable, word or sentence boundaries [Smith, 1997]. Therefore, in order to account for this type of allophonic variation, one would have to examine units larger than words, and model the triggering context in terms of the prosodic organization of the whole utterance, not just the neighboring segments.

This kind of prosodically-conditioned continuous allophonic variation is found in other languages too. In Korean, the first part of an /a/ is breathier when

---

1 The prosodic organization here refers to the prosodic framework proposed in the works such as Selkirk [1984] and Jun [1993] among others.
it follows the aspirated fricative /sʰ/ than when it follows the tense fricative /s*/. This breathiness or aspiration in the following vowel, along with the duration of the fricative itself, is a cue to the contrast between /sʰ/ and /s*/. At the same time, the amount of breathiness after /sʰ/ varies depending on whether the fricative is word-medial or word-initial.

Moreover, there are three types of prosodic grouping that can lend different word-initial positions in the prosodic hierarchy of Korean: prosodic words, accentual phrases, and intonational phrases. (See Section 1.2 for details.) The amount of breathiness in the following /a/ changes depending on which word-initial position the fricative segment occupies. The durational difference between /sʰ/ and /s*/ also differs among these positions. Therefore the nature of this variation and distinction can only be fully understood within the prosodic hierarchy of Korean. If we wish to build a TTS system that sounds natural as well as comprehensible, we have to make it mimic these variations inherent in natural utterances.

To understand how we propose to mimic this phonological allophonic process, it is important to understand how TTS systems have mimicked the effects of coarticulation. Coarticulation refers to continuous variation conditioned by neighboring segments rather than by prosodic position alone. A typical example cited is the articulation of /k/ in the words key, cow, car and so on. The /k/ sounds in these words are different from each other because the /k/ in key and cow are more or less palatalized due to the following high or low front vowel.

The implementation of coarticulation in a TTS system crucially depends on the choice of the synthesis technique. We can classify the different approaches as more or less “knowledge-based” (or more or less “ignorance-based”, to use Jan van Santen’s joking term for the other end of this continuum of approaches). An extensive amount of knowledge about the exact degree of palatalization is required for a “knowledge-based” approach such as a formant-based system, where explicit coarticulation rules are used to specify the fricative pole frequencies in the /k/ burst and the second and third formant locus frequencies. On the other hand, little knowledge is required for an “ignorance-based” approach such as a diphone or demisyllable concatenation system, where the concatenation units containing any given segment are chosen from different segmental contexts in order to contain in themselves the most drastic typical coarticulatory effects of neighboring segments.

We adopted an analogous approach to this in implementing the continuous allophonic variation associated with prosodic context. We chose our concatenative units not just on the basis of the segmental context, but also on the

\(^2\) The fortis series of Korean obstruents will be marked with an asterisk symbol.
basis of the prosodic context — i.e., the position of the units with respect to the edges of constituents that are known or suspected to be the relevant prosodic domains for this type of variation.

We adopted the theory of prosodic domains of Korean laid out in Jun [1993] in the design of our system because we believe that this framework can be used to adequately implement the continuous allophonic variation by mapping each relevant continuum onto the discrete ranges associated with different prosodic domains defined in the framework. We believe that this approach is reasonable, because recent studies on Korean (see Section 1.1) provide increasing evidence that the prosodic categories of this framework act as the domain of application of some of the allophonic as well as allomorphic processes of Korean.

The traditional way to capture allophonic variation was by encoding all segmental variation other than coarticulation as categorical feature changes (e.g., SPE-style allophony rules) using rewrite rules, as in the letter-to-sound rules and the post-lexical rules module of the Festival system [Black et al., 1999]. The phonological component of the MITalk system has six rewrite rules of segmental phonology that would help listeners parse the boundaries of words and phrases. One of these rules replaces /t/ or /d/ with the alveolar flap across word boundaries. That is, if the plosive /t/ is followed by a non-primary-stressed vowel and preceded by a nonnasal sonorant, e.g. *sat on a chair*, it is replaced by an alveolar flap [see Allen et al., 1987, p.87]. Some of these same rules were implemented in the postlexical component of the Festival system [Black and Lenzo, 2003], but as the developers admit, the scope of the component needs further research. The problem is that prosodically-conditioned allophonic variation cannot be specified completely and systematically in this way. For example, while the postlexical rules could specify a categorical alternation between flap and /t/ in *sat on the chair* versus *at all*, or between aspirated [kʰ] in *the car* and *this car* versus unaspirated [k] in *discover*, it would be very unwieldy to use re-write rules to capture the different degrees of aspiration for the variants of /k/ in *this car, this career, and this task*.

Whether a concatenative system uses fixed units such as diphones or variable unit selection, it will not be able to systematically encode continuous prosodically conditioned allophonic variation using postlexical rule specification of categorical alternations such as flap versus [t] or aspirated [kʰ] versus unaspirated [k]. For example, in a diphone synthesis system, a particular diphone unit, once obtained, is used repeatedly wherever it is required, irrespective

---

3 For this type of system, a speech corpus of annotated speech units is necessary. The quality of the synthetic speech partly depends on the length of the speech unit. The longer the speech unit, the more natural the coarticulation of the output speech. For example, if the entire input string matches one of the speech units, the whole unit is used for synthesis with little segmental manipulation compared to a diphone synthesis system.
of its location in a sentence. This is an inevitable consequence of the system design principles when such prosodically conditioned allophonic variation is only poorly understood.

Moreover, the system is unwieldy even when variation is better understood. For example, if we wished to follow the rewrite formalism in building a TTS system, we would have to write rules that involve a feature like [breathy] to represent the vowel /a/ following Korean /sʰ/ and use the breathy vowel whenever it is preceded by the fricative. However, as mentioned above, much of this kind of allophonic variation is fairly continuous.

A more satisfying approach, both computationally and linguistically, to modeling this kind of continuous variation is to encode the context directly in the choice of concatenative unit. Although concatenative TTS systems do not allow the same explicit continuous control of segmental parameters that it provides for controlling duration, F0, and intensity, it is possible to build a system that does capture a great deal of regular prosodically conditioned segmental variation by recording different concatenative units for different prosodic position. For example, by choosing diphones as the concatenative units, the segmental context that triggers breathiness will automatically be encoded in the name of the [sʰ-a] diphone.

Furthermore, since the prosodic phrasing needs to be specified in order to provide the control structures for duration and F0 generation modules (see Yoon [2006] for details of how we built a prosodic phrasing model), the prosodic conditioning of degree of breathiness can also be modeled by recording different diphones for the “same” [sʰ-a] unit in the four different prosodic contexts. With an accurate phrasing model, the prosodically sensitive diphones would be employed in the “correct” prosodic positions.

Not capturing these finer-grained differences could make the synthetic speech sound less natural. Moreover, if the variation is systematically related to the percept of the prosodic structure (as suggested by Fougeron [2001], among others), then not capturing these differences accurately will make the speech less comprehensible. So, what we are proposing to do is to use contentative synthesis to systematically encode as much of the relevant “categorical” information about the prosodic structure directly in the specification of the concatenative unit, so that the system can combine the advantages of concatenative synthesis with the knowledge of prosodic structure that we have from descriptions such as Jun [1993].
1.1 Review of literature

Earlier studies usually focused on word-level environments, e.g. whether the target segment is initial or medial to a word or a syllable, or whether the syllable containing the target segment is stressed or not. As will be shown below, however, there is an increasing number of studies that suggest that such word-level phenomena should be considered within the prosodic hierarchy of the language involved.

In a study with American English /z/, Smith [1997] showed that /z/ is produced with varying amount of devoicing depending on whether the segment is final just to a syllable, or ends a word or is the last in the sentence as a whole. More important was the fact that devoicing occurred more often at the end of the larger domain (sentence) than at the end of a smaller domain (syllable). All four speakers of her production experiment produced all sentence-final /z/’s with complete devoicing, most of their word-final /z/’s in the utterance pause before with partial devoicing, and one or more of their syllable-final (but word-medial) /z/’s in husband with full voicing. Smith suggests that these observations confirm the hypothesis that gestural magnitude reduces more at the ends of larger prosodic domains, the mirror image of the hypothesis that higher prosodic domains accentuate the gestural magnitude of the domain-initial segments.

If the two hypotheses were true, it would imply that the segments initial or final to different levels of the prosodic domains should be treated differently in TTS system development. For example, if the same diphone containing the American /z/ were used at the end of a syllable, a word, or a sentence, it would fail to properly model the allophonic variation observed in Smith’s study. Even if the duration module allows for the durational differences of the fricative and the preceding vowel, the same diphone cannot be used to generate the varying degree of voicing or devoicing. Nor can this problem be resolved by employing an additional postlexical rule that, for example, replaces a sentence-final /z/ with a voiceless /s/, since then the duration rules could not refer to the contrast between /z/ and /s/ to generate the durational differences in the fricative and preceding vowel. Also, the system cannot handle word-final /z/’s that were partially devoiced unless the system adds another postlexical rule that generates a different segment intermediate between /s/ and /z/. The results of Smith’s study clearly favor a TTS approach that is sensitive to the phrasal organization of an utterance. In this respect, designing a diphone database which is sensitive to the prosodic organization of the target language is a consequence of studies such as hers.

In Keating et al. [1998], a study of the domain-initial consonants of the four languages English, French, Taiwanese, and Korean, an attempt was made to
investigate ways in which prosody affects individual speech segments. Their target consonants, unaspirated /t/ and /n/, were embedded in different carrier sentences and produced by two or three native speakers of French, Korean, and Taiwanese. Keating and her colleagues conclude that there is phrasal/prosodic conditioning of articulation across languages. In other words, phonetic properties of individual segments depend on their positions in the prosodic structure.

Pierrehumbert and Talkin [1992] studied the relationship of English /h/ and glottal stop /ʔ/ with respect to the English prosodic structure. Specifically, the materials varied position relative to the word prosody (the location of the word boundary and the word stress) and relative to the phrasal prosody (the location of the phrase boundary and the phrasal stress as reflected in the accentuation). They demonstrated that the pronunciation of /h/ and /ʔ/ depends on word- and phrase-level prosody. Their experiment showed that being in an accented syllable or adjacent to a phrase-initial boundary in English increases gestural magnitude. They also showed that when a vowel-initial word is situated at the beginning of an intonational phrase, the vowel has a much higher likelihood of being glottalized than in other locations, for both stressed and unstressed vowels.

Fougeron [2001] investigated articulatory properties of segments of various types (/t, k, s, l, n/ and /i, ā/). The segments were embedded in initial position of four prosodic constituents in French: the intonational phrase, the accentual phrase, the word and the syllable. She found that the articulation of initial segments varied depending on the prosodic level of the constituent. Target segments at the beginnings of higher prosodic domains had more linguopalatal contact, longer lingual closure as reflected in the EPG signal, longer acoustic duration as measured from the audio signal, less nasal flow for nasals, and more glottalization for vowels. Her results confirmed that for French the articulation of a speech segment is affected by the prosodic organization of the utterance in which it is produced. She concludes that variation in the articulation of initial segments reflects the hierarchical organization of the prosodic constituents and that this variation is not produced randomly. In the conclusion of her study, she claims that the perception of phrasing, frequently said to be cued by such suprasegmental aspects as the final lengthening and melodic contours, is also reflected in the articulation of initial segments, and that these ‘segmental’ properties should be considered as one of the suprasegmental characteristics of speech.

One interesting question that Fougeron asked in her study is whether these differences in the articulatory properties of segments are ‘perceptible’ prosodic markers, in other words, whether listeners use them for prosodic segmentation. Although she could not draw any definitive conclusion about this from her production study, she emphasizes that the articulatory properties of initial segments appear as traces of the prosodic encoding of an utterance. The
exact nature of the ‘decoding’ by the listeners needs to be studied with more perception studies. One of the goals of this work is to explore this question for the specific case of Korean, by building a prototype TTS system that can be used to produce the stimuli for a listening test. In the next section, we show why Korean is a good testbed for this kind of system and listening test.

1.2 The prosodic hierarchy of Korean

Fig. 1. Prosodic hierarchy of Seoul Korean. IP, AP, W, and σ stand for intonational phrase, accentual phrase, prosodic word, and syllable respectively. Courtesy of [Jun, 2000]

The prosodic system of standard Korean differs from that of English and many other languages that are comparably well-studied in that the notions “stress” and “accent” are not particularly useful. That is, it is possible to describe most of the salient prosodic structure of the standard dialect (and of most regional dialects other than Kyongsang) just in terms of the “junctures” or “breaks” that organize strings of consonant and vowel segments into a hierarchy of bracketed groups. The following description of this hierarchy, along with Figure 1, is from Jun [1993, 2000]. The figure is a schematic diagram (a metrical tree) showing the four levels of prosodic grouping that have been established, along with the tonal elements that mark the edges of the top two levels of hierarchy.

The syllable (represented by σ in the diagram) is defined by the four structural positions given in the rewrite rule below, where C, V and G represent consonants, vowels and glides respectively:

\[ \sigma \rightarrow (C)(G)V(C) \]

Above the syllable is the prosodic word (PW or W in the diagram). Above
the level of the PW are two prosodic categories defined by the distribution of various low (L) and high (H) tones in the intonation contour of the utterance. These are the accentual phrase (AP) and the intonational phrase (IP). An IP is marked by pronounced final lengthening as well as a final boundary tone, chosen from the set of L%, H%, LH%, HL%, HLH%, HLHL%, LHLH% and LHLHL%. Each boundary tone is a pragmatic morpheme in its own right, like the boundary tones of English, Japanese, Cantonese, and many other languages.

The full AP tone pattern is a more or less steep rise anchored to the first two syllables of the phrase followed by an interpolated fall to a low tone and subsequent rise on the last two syllables, i.e. a LHLH pattern. Because high pitch is phonologized as a cue to the phonation type contrast among tense, aspirated, and lenis obstruents, and because of constraints on tonal crowding, the surface tonal pattern can differ from this full “underlying” LHLH pattern. First, the initial rise that marks the beginning of an AP is typically much steeper if the initial segment is either aspirated or tense, because the initial low tone target is then displaced by a left-ward spreading of the high tone. Second, in the case of a very short (one or two syllable) AP, the initial rise is deleted, and only the final LH is realized. Finally, in the last AP within an IP, the rise at the end of the AP is replaced with the boundary tone of the IP.

As suggested above, the four levels of grouping diagrammed in Figure 1 are the relevant structures that define the notion of “prosodic context” for a segment. Moreover, there are different types of syllable-initial position. That is, implicit in the diagram is the fact that a syllable-initial consonant might also be initial to its PW. A word-initial consonant, in turn, might also be initial to the AP that contains the PW, a prosodic context where the occurrence of a H tone is one of the cues to the specification of a word-initial obstruent as tense or aspirated rather than lenis. As the diagram shows, then, there are three types of word-initial position (i.e. initial to a PW but AP-medial, initial to an AP but IP-medial, and initial to an IP) and four types of syllable-initial positions (i.e. initial to a PW-medial syllable as well as the three types of word-initial position). These prosodically different initial positions have been the focus of many studies that investigated prosody and its relationship with speech segments.

Jun [Jun, 1993] compared the VOT of Korean /pʰ/ in three positions: initial in an accentual phrase, initial in a prosodic word and medial in a word. The VOT of the segment in higher domains was longer than that in the lower domain. Specifically, the duration of VOT was generally longer phrase-initially than phrase-medially. It was also longer word-initially than word-medially. Based on her result, she concludes that three groupings, i.e. phrase-initial, phrase-medial, and word-medial groups, are significant for the VOT differences of the Korean /pʰ/.
Jun’s results were replicated and extended in a later study by Cho and Keating [2001], who examined the effect of prosodic position on segmental properties of Korean /n, t, tʰ, t*/. They found that consonants initial to higher prosodic domains were articulatorily stronger than those initial to lower domains in the sense that the former had more linguopalatal contact and longer stop seal duration than the latter. Based on this result, they claim that in Korean the domain-initial “strengthening” and “lengthening” is a single effect.

In a study on how the articulatory and acoustic properties of Korean fricatives are affected by prosodic structures, Kim [2001] demonstrated an effect of position in the prosodic hierarchy on the articulation and acoustics of the two Korean fricatives /sʰ/ and /s*/. Her results showed that speakers made a distinction between at least two prosodic domains along several articulatory and acoustic parameters. In another production study on Korean fricatives, Yoon [2005] also showed the effect of prosodic position on the two segments. In addition, he claimed that a simple durational modification cannot adequately model the dynamic change in the ratio of the frication and aspiration noise in the aspirated fricative /sʰ/, which we believe is another indication for the use of prosodically-sensitive diphones.

2 System design

2.1 Prosodic diphones

Our working hypothesis was that the effects of prosodic position are the primary type of allophonic variation that we need to model, and that we can encode that variation more naturally by recording different variant diphones for each different relevant prosodic position. We will call these position-specific diphones “prosodic diphones” as opposed to “conventional diphones” that are only indirectly sensitive to their positions in an utterance if at all. Since it is the domain-initial consonants that were reported to be affected by the prosodic hierarchy and that the right-hand side diphone of the consonants will contain the beginning half of the following vowel, we also assumed that the prosodic positional effects extend halfway to the vowel of the domain-initial or domain-final syllables. The assumption was adopted because the effect of the

4 That is, “conventional diphones” are defined entirely by the features that specify the phoneset. If a particular pattern of prosodically conditioned allophonic variation can be captured by specifying a categorical substitution in the “postlexical rules” module, then diphones involving that set of variants will be prosodically sensitive.

5 This will be problematic for certain cases – most obvious being sequences involving one of the high vowels in a devoicing environment.
consonant phonation type for Korean coronal fricatives extends about halfway to the following vowel in careful speech if the following vowel is the longest /a/ [Yoon, 2003]. More study needs to be done to extend this assumption to the other types of consonants but for now this will be part of our working hypotheses.

A conventional diphone such as p-a in Korean can occur in four different prosodic domains: initial to an IP, to an AP or to a PW and medial to a PW. In order to differentiate these prosodically different versions of the same conventional diphone, we employed three sets of boundary symbols, < and >, [ and ], and { and }, to represent the beginning and end of the three prosodic domains, IP, AP and PW, respectively. The diphones medial to a PW were specified by the absence of any of these symbols.

A diphone with any one of the above boundary symbols will be called an “edge diphone” as opposed to a non-edge diphone medial to a PW. The four prosodic versions of our example diphone thus will be <p-a, [p-a, {p-a and p-a}. This notation represents our hypothesis that a syllable-initial phone p before a will be realized as four different allophones in the prosodic hierarchy of Korean. Likewise, a syllable-final stop b in a diphone a-b is assumed to be realized as either a-b>, a-b], a-b} or a-b. A short sample sentence consisting of one IP, two AP’s, three PW’s and ten syllables will have the following pattern (Periods represent syllable boundaries and bold italicized diphones represent edge diphones):

<jeong.tong.bu} {gwan.gye.ja.neun} [mal.hess.da>
An official from the Ministry of Information and Communication said.
<CV.CVC.CV} {CGV.C.GV.CV.CV} [CV.CCV.CV.V>

We used the phone set for Korean as described in Yoon et al. [2002]. Korean has the syllable structure (C)(G)V(C) and it consists of 7 vowels (V) and 22 consonants (C). These contain 2 glides (G), w and y, and 1 silence phone. In this study, we treated sequences such as CwV and CyV as consisting of a labialized or palatalized consonant plus vowel diphone, i.e. C-w-V rather than as a C-w w-V sequence. Depending on how one segments the diphones from the read corpus, a glide can be treated as part of the preceding consonant or part of the following vowel. We opted for the former. The reason for this choice was that we believed that the glides affect the characteristics of the burst release of preceding consonants and thus if the stop-glide cluster is broken apart, the stop with a hint of glide effect on the burst release could be used in the “wrong” position in an utterance.  

In addition, the assumption was unavoidable because of the technique that we chose, i.e. a diphone-based concatenative synthesis system.

However, with hindsight, we realized that the other option could work as long as the diphone boundary inside the stop-glide cluster is placed with respect to the
consonant symbols. Adding twenty labialized and palatalized consonants, the total number of the members of the Korean phone set was seventy.

When three sets of different boundary symbols were assigned to these conventional diphones, the total number of phones increased seven fold, i.e. to 490 phones in total. When all possible combinations of prosodic diphones and phonotactic constraint of Korean are considered, 6,503 theoretically possible prosodic diphones were obtained; 1,270 IP edge diphones, 2,075 AP edge diphones, 2,075 PW edge diphones and 1,083 PW internal diphones.

2.2 Corpora

In order to obtain as many of these theoretically possible diphones as possible, we had two options. One was to come up with a nonsense carrier sentence to which we embed all prosodic diphones divided into sets involving different segments, and the other was to use real speech corpora. We could use the recording of “real” words as in Fitt [2001] or the recording of the actual utterances. We decided to choose the second option because we had a read version of a subset of the Korean Newswire Text Corpora [Linguistic Data Consortium, LDC] available to us. This approach could also test one of our questions, i.e. whether a real speech corpus can be used in a diphone-based concatenative TTS system. If successful, this implies that existing K-ToBI labeled corpora could be used to build a prosodically sensitive diphone database.

end of the stop gap so that the left-half of the stop-glide cluster does not include any of the stop burst energy and that the right-half of it contains all of the burst energy. This way, the left-hand side of the cluster will not have any hint of the glide, because the silent stop gap does not contain any spectral energy of the burst. With this approach the glides can be freed from the stops and can now be treated as part of the diphthongs, thus making the vowel inventory include diphthongs. This can “un-triple” the number of consonant symbols, but double the number of vowel symbols.

Of course, this task involves the use of a talented professional who is familiar with the prosody of Korean. If this is to be done by a naive speaker, another approach should be taken, for example, devising sentences that will naturally invoke intended phrasings. This nonsense carrier sentence should not be unreasonably long, and at the same time has to have all the prosodic “slots” that a particular segment can occupy in an utterance. For example, according to our “prosodic diphones” view, a segment such as /b/ can be placed initial/final to an IP, to an AP, and to a PW, and medial to a PW. If the carrier sentence can hold all these prosodic positional “slots”, then we can embed all the consonants one by one into the prosodic slot with consideration of the vowel context. A preliminary calculation shows that 567 repetition of the carrier sentence with different segments embedded can cover all prosodic diphones for Korean. The actual number of repetitions can be much fewer than this if the palatalized/labialized consonants are treated differently.
A 400-sentence subset of the Korean Newswire Text Corpus [Linguistic Data Consortium, LDC] was read by a male native speaker of Seoul Korean (the first author). The corpus contained 28,666 syllables or 9,246 words, with an average of 72 syllables or 23 words per sentence. The histograms in Figures 2 and 3 illustrate the distribution of the number of words and syllables of the corpus.

Fig. 2. Histogram of the number of words per sentence for the Korean Newswire Text Corpus

Fig. 3. Histogram of the number of syllables per sentence for the Korean Newswire Text Corpus

The Korean Newswire text corpus was read by the speaker because we intended to build a diphone-based concatenative TTS system for the reading style of the speakers. The text of the speech corpus was morpho-syntactically
parsed following the conventions of the Penn Korean Treebank project [Han et al., 2002] and its read version was prosodically labeled following the K-ToBI (Korean Tones and Break Indices) prosodic labeling conventions [Jun, 2000]. The parsing and the labeling was done by a trained labeler (the first author).

In Korean, as in many other languages, the orthographic form of a word is different from its pronounced form. However, a systematic relationship exists between the two. The relationship between the phonology and the orthography of a language depends on the type of writing system. In a “phonographic” writing system such as the one used for Finnish, where the words are written as they are pronounced, the morphophonology of the language is reflected in the orthography and no morphophonological rules would be required to apply to get the pronounced form of words. In a “morphographic” writing system such as Korean hangul, where the identity of the base form of a word is preserved in all of its inflected/derived forms, the morphophonological rules are not reflected in its orthography and grapheme-to-phoneme (GTP) rules can be applied to get the reasonably accurate form of pronunciation. We followed Yoon and Brew [(In Press] and obtained pronounced forms of the text corpus. The pronounced forms of the text corpus were inserted to the “orthographic word” tier.

The unit word was defined as follows. Since Korean lexical items can be inflected with prefixes, suffixes, postpositions, tense morphemes, etc., by word we mean a fully inflected lexical item, an eojeol. An eojeol corresponds to a space-delimited orthographic word unit.

The prosodic labeling proceeded in three steps. First, each of the recorded sentence was loaded into Praat [Boersma, 2005] and manually segmented by the “orthographic word”. That is, each eojeol was assigned to an interval in the “words” tier, and then the labeler shifted interval boundaries between eojeol to align them to the edge segments in the acoustic signal and to correct any errors in the output from the grapheme-to-phoneme (GTP) module.

In the second step, the phonological tone tier and phonetic tone tier as well as the break index tier were labeled by the labeler with the help of Praat scripts. The default tone labels were assigned to the phonological and phonetic tone tiers, which were manually corrected by the labeler. Assuming that each eojeol corresponds to an accentual phrase, the label ‘LHa’ was assigned to each AP. The label ‘LHa’ is used in the K-ToBI framework to refer to an AP boundary in the phonology labeling tier because it is assumed that the underlying tonal contour of an AP is a rising tone from a low tone (L) to a high tone (H). The ‘a’ in the label is there to distinguish the AP from an intonational phrase (IP) label such as ‘HL%’, which means the boundary tone of the IP is also a rising tone. The ‘%’ signals that it is an IP label. The script also assigned a default break index of 2 to each AP and a break index of 3 and a HL% boundary tone
at the end of each sentence.

After the manual correction in the phonology labeling tier, another script assigned default tone labels to the phonetic tone tier. The surface tonal representation of an AP in the phonetics labeling tier can vary depending on the number of syllables and the identity of the phrase-initial consonant. A LH phrasal tone was assigned to an AP with two syllables, a LLH phrasal tone to a three-syllable long AP and a LHLH phrasal tone to an AP which was four or more syllables long. When the initial consonant segment was either aspirated or tense, the first low tone was replaced with a high tone, i.e. HH, HLH or HHLH tones were assigned. The labeler then manually corrected these default labels. The labeled corpora, which we call OSU TalkBank, consist of time-aligned *eojeol*, AP and IP boundaries, etc. as produced by the speaker.

The labeled corpora, which are the first parts of what we call “OSU Korean TalkBank”, consist of the read version of the 400 subset of the Newswire Text corpus, prosodically labeled following the Korean ToBI conventions [Jun, 2000]. Note that we segmented the corpus by the prosodic word not by the phone. The labeling and segmentation of the phones, hence diphones, was done after we selected the minimal number of diphones at a later stage that would cover all 400 sentences.

The component prosodic diphones for each sentence were extracted by another Praat script. The script extracted prosodic diphones based on the segmental compositions and the kind of prosodic domains present in each utterance and added another labeling tier consisting solely of extracted prosodic diphones. Given the GTP-converted example sentence in (a) consisting of one IP and two APs, 21 prosodic diphones were extracted by the script as shown in (b).

(a) ♯< b o g o s eo n eu n] [s eo l my eo ng h e d dd a> ♯

*The report said*

(b) ♯<-<b, <b-o, o-g, g-o, o-s, s-eo, eo-n, n-eu, eu-n], n]-[s, s-eo, eo-l, l-my, my-eo, eo-ng, ng-h, h-e, e-d, d-dd, dd-a>, a>-♯

With the prosodically labeled utterances and their component prosodic diphones, the four hundred sentences were ready for a “greedy” search for the minimal number of sentences that contain all the prosodic diphones that can synthesize all four hundred utterances.
2.3 Diphone selection

The greedy algorithm [van Santen, 1992] implemented in a Praat script was applied to the component prosodic diphones tier of each of the 400 sentences to search for the minimal number of sentences that would cover all the prosodic diphones present in the corpus. In each loop of the script, the prosodic diphones tier of each sentence was checked for the number of distinct prosodic diphone types and the sentence with the most types was chosen. After the selection, the same diphone types as the ones contained in the selected sentence were removed from the rest of the sentences before the subsequent loop began. The algorithm did not take the quality of the diphones into account. Thus, the coverage statistics in this section should be considered with care.

The search ended selecting 223 out of 400 sentences. The prosodic diphones tier of the 223 sentences now had only those diphones, 2,368 in total (36.4% of the 6,503 possible diphones, in other words, the other 63.6% did not occur in our corpus), that could synthesize all 400 sentences. Figure 4 shows the number of available diphone tokens for each type of all possible prosodic diphones. The quality of the available diphones varied. Therefore, depending on the quality of the diphone to be used in the prototype synthesis system, the quality of the synthetic utterance can also change. The first bar in the histogram represents 4,135 diphone types that our diphone database cannot cover and thus it has zero diphone tokens. The second, third, and fourth bar represent the number of diphone types whose number of available tokens in the selected sentences were one, two, and three respectively. The last bar represents the diphone n-eu which occurred in the diphone database 649 times. When the available tokens were more than one, the first one that the greedy algorithm chose was used, after which followed the spectrographic examination to manually check the quality of the token.

Figure 5 shows the number and percent value of each type of prosodic diphones present in database (shaded) relative to its corresponding “possible” prosodic diphones. In other words, it shows how much of the 6,503 diphones our 400-sentence speech corpus was able to cover.

In Figure 6, the database prosodic diphones (shaded) and “possible” prosodic diphones are compared by the diphone type, i.e. whether the diphone is initial or final to an edge, or crosses the edge.

Figures 7 – 10 show the composition of the database diphones (shaded) and “possible” diphones, i.e. whether the diphone starts with a consonant (C) or a vowel (V). As for the CV composition of the database diphones, more edge diphones start with a consonant than with a vowel and more edge diphones end with a vowel than with a consonant, although in terms of the possible
Fig. 4. Histogram of the number of available diphone tokens or candidates for each type of prosodic diphones. Note the logarithmic scale of the vertical axis.

Fig. 5. Histogram of the four types of the prosodic diphones present in database relative to their matching “possible” diphones.

diphones, more edge diphones start with a vowel. Cross-edge AP diphones of the types C]-[C and V]-[C (C}-[C and V}-[C for cross-edge PW diphones) are greater in number than the other cross-edge diphone types.

These figures tell us that the 400-sentence corpus was not sufficient to cover all possible prosodic diphones of Korean. They also suggest that in order to make our system synthesize arbitrary sentences, we need to come up with a strategy to deal with missing diphones.
Fig. 6. Histogram of comparison of the database edge diphones and “possible” edge diphones with respect to their diphone type. Numbers in parentheses show the percentage with respect to each type. “D” represents a member of a particular diphone.

Fig. 7. Histogram of comparison of the database IP edge diphones and “possible” IP edge diphones with respect to their phone composition. C and V stand for consonant and vowel respectively.

2.4 Diphone segmentation and labeling

The selected prosodic diphones from the greedy search were segmented and labeled in Praat. A Praat script prompted the labeler with each of the 223 sentences containing the selected prosodic diphones and saved the labeled file as an Emu label file appropriate for Festival.

The following guidelines were observed for a consistent labeling of the prosodic diphones. The diphone boundary for non-glide monophthongs was given at the most stable region with the least amount of formant change, which usually
Fig. 8. Histogram of comparison of the database AP edge diphones and “possible” AP edge diphones with respect to their phone composition.

Fig. 9. Histogram of comparison of the database PW edge diphones and “possible” PW edge diphones with respect to their phone composition.

Fig. 10. Histogram of comparison of the database PW internal diphones and “possible” PW internal diphones with respect to their phone composition.
coincided with the middle of the vowel. The diphone boundaries for the two glides /y/ and /w/ were given at the mid point of the changing second formant. This was not an ideal solution but was unavoidable due to our decision on how to treat the vowels preceded by either of the glides. The situation could be improved by a different approach (See footnote 6 in Section 2.1.).

Stop consonants were divided with respect to the end of the stop gap. This was consistent for both voiceless and voiced stops. Thus the left half of the stop phone contained either the voiced or voiceless stop gap whereas the right half of the phone contained an optional acoustic burst release followed by an optional aspirated noise and the formant transition to the following vowel, usually two vowel pulses. When a stop consonant was not released, which is very frequent in Korean since no word-final stop is released in isolation, the diphone boundary was given at the end of the vowel formant offset of the preceding vowel.

Fricative consonants were divided with respect to the end of the frication noise. The left half of the fricative thus contained only the frication noise while the right half contained the aspiration noise and the formant transition to the following vowel. When the fricative consonants were realized as one of the lax series of stops or nasals, which is possible due to the morphophonology of Korean, they were segmented as such.

The diphone boundary of the affricate consonants was given with respect to the acoustic burst release of the stop portion of the consonants. This applied to both voiceless and voiced affricates. The left half of the affricate thus contained either a voiced or voiceless stop gap while the right half contained the acoustic burst release, the frication and aspiration noise and the formant transition to the following vowel.

Nasal consonants were given diphone boundaries at the end of the nasal murmur, the left half of the nasal consonant containing the whole acoustic nasal murmur and the right half containing the formant transition to the following vowel as well as the acoustic release.

Liquid consonants were labeled in two ways. When a liquid was realized as a flap /r/, it was labeled as if it were a stop, its diphone boundary being drawn at the end of the stop gap. When the liquid was realized as a non-flap sound /l/, it showed spectrographic characteristics similar to those of a nasal, in which case it was labeled according to the criteria defined for nasals.

These diphone labeling guidelines were followed throughout the labeling. However, there were cases where the diphones selected by the greedy search were not of good enough quality. This happened very often for some vowels, especially high vowels, which are often undershot in read speech as well as in spontaneous speech. The glottal fricative /h/ frequently took the formant
characteristics of the following vowel without a distinct frication noise and sometimes dropped completely without an apparent trace. For the former case, a triphone approach could alleviate the problem while for the latter case, a modification to the GTP component would be necessary. Nasal consonants sometimes had little nasal murmur. When this occurred, other diphone candidates were searched and substituted by manual examination. Specifically, all the sentences containing the target diphone was examined one by one until a good substitute was found.

2.5 Diphone database creation

Pitchmarks and LPC coefficients were extracted from the waveforms of the 223 sentences containing the labeled and segmented prosodic diphones as explained in the Festival manual (version 2.0). A voice was created in Festival with the LPC coefficients, and sample sentences were synthesized to test the diphone database. These sample sentences were the remaining 177 sentences that were not selected in the greedy search for the prosodic diphones. No fundamental frequency contour was given at this stage.

3 Evaluation

In this section, a listening experiment is reported designed to test the quality of the diphone database. The working hypothesis for the creation of a prosodically-conditioned diphone database was that segments are affected by the prosody and that prosodically appropriate versions of the same diphone should be used in the synthesis of an utterance. For reasons that follow in the preparation of the experimental stimuli, we chose to compare two types of prosodic diphones: AP-edge diphones versus PW-internal diphones and PW-edge diphones versus PW-internal diphones.

We created IP-sized test phrases. The reason we chose IP’s as the test unit was 1) they were usually surrounded by pauses, thus acting as short “utterances”, 2) they were relatively small, usually containing from 2 to 8 AP’s in the corpus that we used, and 3) we could use the edge diphones that connect AP or PW boundaries as target sites and create two versions of the same test phrase where in the first “good” version, the inter-AP or inter-PW edge diphone is from the “correct” diphone category whereas in the “bad” version, the inter-AP or inter-PW edge diphone is from the “wrong” PW-internal diphone.

With an IP containing, for example, two AP’s, we were able to test the quality of the diphones in one target site, i.e. the diphone connecting the two compo-
nent AP boundaries. The “good” version of the test IP had all four types of the prosodically “correct” diphones, i.e. IP-edge, AP-edge, PW-edge, and PW-internal diphones, whereas the “bad” version had only two types of prosodic diphones, i.e. IP-edge and PW-internal diphones. The bad version was composed of PW-internal diphones except for the two IP-edge diphones at the beginning and end of the phrase. We had to use the two IP-edge diphones because otherwise we could not synthesize the pause surrounding the phrase. We also synthesized phrases containing two PW’s. Then, the only difference between the two stimuli was that in the “bad” version of the stimuli, we replaced the AP-edge and PW-edge diphones with PW-internal diphones, allowing us to compare perceptible differences between i) the “good” AP-edge diphone (hence, “good-AP”) and its matching PW-internal diphone in the “bad” version (“bad-AP”) and ii) the “good” PW-edge diphone (“good-PW”) and its PW-internal diphone in the “bad” version (“bad-PW”). In other words, the “bad” version had PW-internal diphones where AP-edge and PW-edge diphones were intended.

3.1 Methods

3.1.1 Stimuli

As the candidate test phrases, we used the 177 sentences that were not used in the diphone selection procedure. Otherwise, the test phrases may contain some of the diphones originated from themselves. The 177 sentences had 634 IP’s in total (an average of 3.6 IP’s per sentence). Since our diphone inventory does not cover all possible diphone combinations of Korean and the greedy algorithm searched for the least number of sentences to cover the 400 sentences, which were “good” version, our diphone inventory lacked diphones that were necessary in building some of the “bad” versions. We also looked for phrases containing only two prosodic words, in which case we can embed target consonants in between the boundaries of them. In total 96 stimuli were synthesized (24 good-AP + 24 bad-AP + 24 good-PW + 24 bad-PW). 24 target consonants are given in Table 1. All 96 stimuli were composed of either two AP’s or two PW’s. Thus they all contained one target site, where the target consonants were embedded.

Since we do not have modules for assigning durations, F0 contours, and intensity contours, we copied them from the original natural utterances. F0 contours and durations were copied using the PSOLA algorithm implemented in Praat. Intensity contours were also borrowed from the original utterances. In order to copy the three parameters, durations were adjusted first. A synthetic test phrase, composed only of diphones produced by the ‘SayPhones’ function in Festival, was first aligned segment by segment against its natural version by
Table 1
Type and number of target segments embedded in the internal boundary of the test phrases.

<table>
<thead>
<tr>
<th>target segment</th>
<th>number of segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>/b/</td>
<td>1</td>
</tr>
<tr>
<td>/d/</td>
<td>6</td>
</tr>
<tr>
<td>/g/</td>
<td>4</td>
</tr>
<tr>
<td>/p/</td>
<td>1</td>
</tr>
<tr>
<td>/t/</td>
<td>2</td>
</tr>
<tr>
<td>/dd/</td>
<td>2</td>
</tr>
<tr>
<td>/s/</td>
<td>3</td>
</tr>
<tr>
<td>/j/</td>
<td>4</td>
</tr>
<tr>
<td>/c/</td>
<td>1</td>
</tr>
</tbody>
</table>

Creating labeling files, i.e. TextGrid in Praat [Boersma, 2005], for each phrase. Then the PSOLA algorithm [Moulines and Charpentier, 1990] implemented in Praat adjusted the duration of the synthetic test phrase so that it matched exactly with its natural version in terms of their segmental duration.

With the duration matched, the F0 contour from the natural phrase was imposed on the synthetic version using the PSOLA algorithm. The intensity contour from the natural version was next copied to the synthetic test phrase. All these procedures were done by a Praat script except for the segment alignment step, which was done manually. The alignment step required the most precision in the overall parameter-copying procedure. Conventional acoustic landmarks were used to identify corresponding speech segments on spectrograms.

3.1.2 Participants

80 listeners, 37 females and 43 males, participated in the experiment. They were all native speakers of Korean with normal hearing who grew up in Korea at least until the age of 18. Their average age was 30.6.

3.1.3 Procedures

We performed two types of tasks in three sessions. The dictation task was done in the first session and the naturalness evaluation was done in the two subsequent tasks, i.e. rating and preference task. Half of the listeners did the dictation session first and then the other two evaluation sessions, and the other
half did the two types of tasks in the opposite order.

**Dictation task**

In the dictation session designed to test the overall intelligibility of the synthetic stimuli, we presented 24 pairs of “good” and “bad” stimuli and asked the listeners to write down what they heard in hangul. They were allowed to listen to the stimuli as many times as necessary to be able to write the whole phrase. Here we wanted to test explicitly whether the listeners understood or not by seeing whether their transcription was correct. The number of times listeners hit the play button was a measure of how difficult it was to comprehend the synthetic phrases.

**Rating task**

In the rating session, we played 24 stimuli (consisting of pairs of “good” and “bad” stimuli) back to back three times – with the first constituting the reference stimulus, and the second constituting the test stimulus – and asked the listeners to rate the test stimulus with respect to the reference. In half of the trials presented to each subject, the reference stimulus was a “good” version of a synthesized phrase, and the test stimulus was the matching “bad” version, and in the other half of the trials, the order was the opposite. The participants were divided into two groups, with order counter-balanced, so that each subject was presented only once. Each participant was asked to rate the test stimulus in percentage, with the reference stimulus being a 100%. They were given multiple choices from 50% to 150% in the increment of 10%. If the participant heard a test stimulus and judged it to be better than the reference stimulus by 10%, she chose 110%.

**Preference task**

In the third session, listeners were given a hangul script that listed the 24 pairs of test stimuli (“good” + “bad” version, counter-balanced in order). The first stimulus of each pair was numbered 1 and the second 2. The target diphone sites, i.e. the inter-AP or inter-PW edge diphones, were marked on the hangul script given to the listeners. Double quotes were placed around the syllables containing the target diphones and in an additional column, the target consonants were given. Listeners were asked to choose either of the stimulus of each pair. By doing this, we intentionally asked the listeners to focus on the target consonants of those syllables. They were allowed to play the stimuli more than once and had to choose one from the pair of test phrases whose target consonant sounded more natural.
3.1.4 Spectrographic comparison of sample stimuli

Fig. 11. An utterance containing /b/. `<g eu eu i] [b a r e o n eu n>`. ‘his remarks’. (a) natural (b) synthesized with prosodic diphones (c) synthesized with PW-internal diphones. Arrow indicates the location of /b/.

In Figure 11, we compare the “good” and “bad” version of the stimuli containing the target segment /b/ with their matching natural phrase. All the phrases had two component AP’s. The top panel designated as (a) shows the natural phrase, the middle panel (b) shows the “good” version of the synthetic stimulus, and the bottom panel (c) shows the matching “bad” version. The arrow in the spectrogram of the natural utterances indicate the location of the
Regarding the release burst, the lax stop in the “good” version has clear burst energy whereas in the “bad” version, little burst energy is observed.

3.2 Result

Table 2 gives statistical analyses of the data by items and by subjects respectively. The statistical analyses were performed using GNU R. The analyses were based on three factor ANOVAs.

- **Factor I:** appropriateness (“good” versus “bad”).
- **Factor II:** break level (“AP” versus “PW”).
- **Factor III:** consonant type (“lenis” versus “not lenis”).

The mean values for each of the tasks including the standard deviation are reported in Tables 3, 4, 5 and 6.

Table 2

<table>
<thead>
<tr>
<th></th>
<th>By items</th>
<th></th>
<th>By subjects</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>play count</td>
<td># of eojol correct</td>
<td>rating</td>
<td>choice</td>
</tr>
<tr>
<td>Factor I (appropriateness)</td>
<td>F(1,88)=7.50</td>
<td>p&lt;0.01</td>
<td>F(1,88)=36.31</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Factor II (break level)</td>
<td>F(1,632)=20.01</td>
<td>p&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor III (consonant type)</td>
<td>F(1,88)=15.11</td>
<td>p&lt;0.001 (I×III)</td>
<td>F(1,632)=14.28</td>
<td>p&lt;0.001 (I×III)</td>
</tr>
</tbody>
</table>

The purpose of the analyses is to examine the effects of the three factors on the comprehensibility of the stimuli used in the dictation and the preference task. Specifically, for the factor I (appropriateness), we tested the effect of using a “good”, i.e. prosodically appropriate, diphone versus using a “bad”, prosodically inappropriate diphone for the target diphone in between the two phrases of the stimuli. In the two-session preference tasks, we tested the effect of using a pair of stimuli in opposite orders, i.e. a “good” stimulus followed by its matching “bad” stimulus or vice versa. For the factor II (break level), we wanted to test which will have more adverse effects on the comprehensibility of the stimuli, replacing the AP-edge diphone with a PW-internal one or replacing the PW-edge diphone with a PW-internal one. For the factor III (consonant type), the effect of segment type was tested.
3.2.1 Dictation task

As shown in Table 2, there was a main effect of factor I (appropriateness), on the mean play counts of the dictation task. In other words, listeners had to play the “bad” stimuli more before they wrote down what they heard. There were main effects of factor I (appropriateness) and factor II (break level) on the mean number of *eojeol* correct for the by-subjects analysis. That is, listeners were more accurate with “good” stimuli and the effect was bigger for AP stimuli. There were no interactions among the factors.

<table>
<thead>
<tr>
<th></th>
<th>good</th>
<th>bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lenis</td>
<td>2.7 (3.2)</td>
<td>3.6 (3.5)</td>
</tr>
<tr>
<td>not lenis</td>
<td>2.7 (2.2)</td>
<td>3.6 (3.8)</td>
</tr>
<tr>
<td>PW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lenis</td>
<td>2.6 (3.2)</td>
<td>3.1 (3.2)</td>
</tr>
<tr>
<td>not lenis</td>
<td>2.7 (2.5)</td>
<td>3.7 (4.3)</td>
</tr>
</tbody>
</table>

Table 3
Mean play count for the dictation task. Numbers in parentheses are standard deviations.

<table>
<thead>
<tr>
<th></th>
<th>good</th>
<th>bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lenis</td>
<td>1.5 (0.6)</td>
<td>1.3 (0.7)</td>
</tr>
<tr>
<td>not lenis</td>
<td>1.6 (1.0)</td>
<td>1.5 (1.1)</td>
</tr>
<tr>
<td>PW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lenis</td>
<td>1.7 (0.6)</td>
<td>1.6 (0.7)</td>
</tr>
<tr>
<td>not lenis</td>
<td>1.7 (0.8)</td>
<td>1.5 (0.8)</td>
</tr>
</tbody>
</table>

Table 4
Mean number of *eojeol* correct for the dictation task. Numbers in parentheses are standard deviations.

3.2.2 Rating task

In the rating task, there was a main effect of factor I (appropriateness) on the mean percent rating, which means that the listeners liked the good stimulus of each pair. There was an interaction between the factors I (appropriateness) and III (consonant type) in the by-items analysis. There was an additional interaction between the factors I (appropriateness) and II (break level) in the by-subjects analysis. As Figure 12 shows, the main effect was greater for the lenis and AP stimuli. The uneven coverage of plosives in the test set could have affected the observed interaction.
Fig. 12. Interaction for the mean percent rating. Left: Factors I (appropriateness) × III (consonant type), Right: Factors I (appropriateness) × II (break level).

<table>
<thead>
<tr>
<th>Rating 2nd stimulus relative to 1st</th>
<th>good-bad</th>
<th>bad-good</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>lenis</td>
<td>96 (15)</td>
</tr>
<tr>
<td></td>
<td>not lenis</td>
<td>98 (13)</td>
</tr>
<tr>
<td>PW</td>
<td>lenis</td>
<td>97 (15)</td>
</tr>
<tr>
<td></td>
<td>not lenis</td>
<td>101 (12)</td>
</tr>
</tbody>
</table>

Table 5
Mean percent rating for the preference task. Numbers in parentheses are standard deviations. ‘good-bad’ means that the first stimulus was good while the second was bad.

3.2.3 Preference task

Fig. 13. Interaction for the mean choice. Left: Factors I (appropriateness) × III (consonant type), Right: Factors I (appropriateness) × II (break level).

There was a main effect of factor I (appropriateness) on the mean choice. That
Table 6
Mean choice for the preference task. Numbers in parentheses are standard deviations. ‘good-bad’ means that the first stimulus was good and the second was bad. The first stimuli were given number 1 and the second number 2 in the response sheet.

<table>
<thead>
<tr>
<th></th>
<th>lenis</th>
<th>not lenis</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>1.3 (0.5)</td>
<td>1.4 (0.5)</td>
</tr>
<tr>
<td>PW</td>
<td>1.4 (0.5)</td>
<td>1.5 (0.5)</td>
</tr>
</tbody>
</table>

is, listeners liked the “good” stimulus of each pair better than the other. As in the rating task, there were interactions between factors I (appropriateness) and II (break level) and factors I (appropriateness) and III (consonant type). Figure 13 shows that the main effect was greater for the AP and lenis type.

4 Discussion

This paper presented a prosodically conditioned diphone database created from a prosodically labeled read speech corpus and a listening experiment designed to test the database. Our hypothesis was that the effects of prosodic position on sound segments are the primary type of allophonic variation that we need to model. We encoded that variation by recording different variants for a particular diphone from different relevant prosodic positions. We also hypothesized that the prosodic positional effects extend half-way to the vowel of the prosodic domain-initial or domain-final syllables.

One of the questions that we asked at the beginning of this work was whether it would be plausible to use “real” utterances in the construction of a diphone database. Our listening test suggests that it is a viable option. We can now claim that not only “real” words [Fitt, 2001] but also real sentences can be used for recording diphones. Our approach is potentially better than that of hers because whereas she took the diphones from the middle of words, our diphones are sensitive to the prosodic positions.

A problem of our approach is that it cannot cover an unlimited number of sentences. This was expected from the beginning when we decided to use a 400-utterance read corpus. However large a real speech corpora we may use, we would eventually suffer from the diphone coverage problem. One way to overcome this problem which we did not implement in our work here would
be to analyze the type of missing diphones and replace them with existing
diphones. A missing diphone with a stop-stop sequence can be replaced with
another stop-stop diphone that exists in the diphone database. The more the
segment type matches with respect to its phonetic features, the less the flaw
of the synthesized sentence will be noticeable.

An alternative to using real speech corpora for diphone database creation
would be to use nonsense carrier sentences designed to simulate required
prosody (See the first footnote of Section 2.2). A carefully chosen “prosody-
carrier” nonsense sentence could embed a set of prosodic diphones. Based on
our estimation, a 567-sentence corpus generated by such nonsense prosody-
carrier sentences can cover all possible prosodic diphones in Korean. Studies
such as Jun and Oh [1996] and Jun et al. [1997] suggest that it is possible to
impose necessary prosodic domains to a prosody-carrier sentence.

Although our work started as a limited domain synthesis in the sense that
our diphone database only covers the 36.4% of all possible prosodic diphones
of Korean and that it can only synthesize the 400-sentence corpus and other
sentences that do not contain the diphones that our system lacks, its outcome
strongly suggests the plausibility of using K-ToBI labeled speech corpora in
the diphone database creation.

The result of the listening experiment confirmed our initial working hypothesis
that the effects of prosodic position are reflected in the variation of speech
segments. It also demonstrated the fact that listeners are “sensitive” to the
prosodic effects reflected on the speech segments. When presented with stimuli
synthesized from the “good”, prosodically appropriate diphones, listeners were
able to do the dictation quickly and liked them better in the preference task.
This is evidence that prosodically appropriate diphones should be used in the
synthesis of an utterance.

In terms of the durations of the segments used in the synthesis of our stimuli,
note that we copied the durations of our synthetic stimuli from their original
natural utterances. The fact that the synthetic speech segments had “perfect”
durations and that listeners still preferred our prosodically conditioned seg-
ments is evidence, as Yoon [2003] claimed, that the durational correlate of
the prosody cannot be properly modeled just by manipulating the segmental
durations of the component diphones of the synthesized phrases.

Our work has implications for other types of synthesis techniques such as
the large inventory unit selection synthesis. Unit selection synthesis technique
annotates each unit with large amounts of lexical context, which implicitly
provides information about the prosodic status of a given segment. However,
the result of our experiment suggests that the technique could still benefit from
the explicit inclusion of prosodic information specified in the ToBI prosodic
labeling conventions.

Note also that the result represents the best-case situation because we copied the durations, phrasing and F0 contour from the original natural utterances. In the more realistic situation where the duration and F0 are also generated by the TTS synthesis system, people would be less likely to notice segmental problems.

5 Conclusion

This paper examined the quality of the prosodically sensitive diphone database, by reporting the listening experiment performed on eighty native speakers of Korean with the stimuli synthesized from the diphones. The listening experiment showed that prosodically appropriate diphones yield better synthetic utterances. It also supported the claim [Yoon, 2005] that simple manipulation of durations cannot properly model prosodic effects on segments. We interpret the result of our listening experiment as supporting the view that prosody has segmental correlates as well as tonal correlates and that listeners perceive the segmental encoding of prosodic domain information.

References


Chunghye Han, Eon-Suk Ko, Heejong Yi, and Martha Palmer. Penn Korean Treebank: Development and evaluatio. In *Proceedings of the 16th Pacific As-


Kyuchul Yoon, Chris Brew, and Mary Beckman. Letter-to-sound rules for