Phonetics of Segmental F0 and Machine Recognition of Korean Speech

Tae-Yeoub Jang

Thesis submitted for the degree of Doctor of Philosophy
University of Edinburgh
2000

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Tae-Yeoub Jang

Abstract

The main goal of the study is to improve performance of Korean automatic speech recognition by exploiting the fundamental frequency (F0) of vowels, which is affected by identity of the preceding consonant. The hypothesis is that if the vowel F0 is given, the identification of the consonant can be more accurate.

The effect, which I will call the "segmental F0 effect", has been confirmed by a number of phonetic studies across various languages. Most frequently, the F0 value of a vowel has been suggested to be a cue to the voiced/voiceless distinction of the preceding consonant. In Korean, segmental F0 can be useful for differentiating the three typical manners (lax, tense, and aspirated) of stop and affricate articulation. Earlier phonetic studies have found that F0 of a vowel onset becomes higher after strong stops (eg., tense and aspirated sounds) and lower after lax stops. It is also suggested that this effect is more salient in Korean than European languages like English and French.

If the segmental F0 effect is going to be helpful for speech recognition, it has to be detectable outside the carefully controlled data used for phonetic studies. I show that automatic measurements over a large amount of data can also capture the effect. Other related issues regarding segmental perturbation which have not been dealt with in earlier studies are also investigated.

Integration of the segmental F0 effect with speech recognition is achieved using demisyllables as basic recognition units. As some demisyllables are composed of both an onset consonant and the front part of the nucleus, it is relatively easy for them to carry characteristics of the consonant-vowel relation, such as segmental F0, on their own. Besides, I find that an HMM demisyllable based recogniser performs better than a baseline HMM recogniser with phone-like units even before F0 is included. Thus, using demisyllables in Korean speech recognition has an independent motivation. In addition, a lexicon modification technique by pronunciation modelling is introduced to further enhance the recognition performance. I show that inclusion of F0 in the demisyllable recogniser gives further improvement in results.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

This is a study of segmental perturbation of fundamental frequency (F0), and its application to automatic speech recognition (ASR) of Korean speech. The research is focused on the way vowel F0 is influenced by the manner of articulation of a preceding consonant, especially by a stop or affricate in prosodic word initial position. The main goal is to improve automatic speech recognition performance exploiting segmental perturbation of F0. As a preliminary, replication and extension of relevant acoustic phonetic studies, originally carried out on small sets of hand-labelled data, will be conducted using automatic analyses of a large amount of data. Though a continuous speech database is mainly used in recognition experiments, isolated words are also employed for phonetic verification. As well as exploiting segmental F0, other enhancements are made to the baseline speech recogniser through lexicon modification using pronunciation modelling and the use of demisyllable units, resulting in improved performance. The success of the current work is especially significant in demonstrating that cues found by independent phonetic or linguistic studies can be exploited for speech recognition.

1.1 The Segmental F0 Effect

There are many factors which affect the F0 contour of an utterance. Among them, the main interest of this study is on local F0 perturbation at the segmental level. Two effects are generally agreed to exist at this level. Firstly, F0 varies depending on the identity of vowels such that high vowels have higher F0 than low vowels. Secondly, consonant type affects the F0 of the following vowel in the same syllable. The main interest in this study is in the latter phenomenon and I will call this the "segmental F0 effect". The alternative terms "microprosody" (Hirst 1983) or "microintonation" (Willems 1982) are avoided lest the expression 'micro' should, as Silverman (1987) warned, give the negative impression that segmental perturbation is much less important than sentence level intonation.

A number of phonetic studies have confirmed segmental F0 effects across languages. In European languages including English, different F0 of vowels is triggered by voicing differences of preceding consonants. The general agreement is that voiceless consonants induce higher F0 whereas voiced consonants induce lower F0 in the following vowel (House & Fairbanks 1953, Lehiste & Peterson 1961, Haggard *et al.* 1970, Mohr 1971, Löfqvist 1975, Hombert 1978, Umeda 1981, Ohde 1984, Silverman 1986, Terken 1995), though there are also disagreements among them on detailed characteristics of segmental F0, such as size of effect, duration of effect, and shape of affected F0 contours.

Korean has a three way contrast of stops and affricates rather than the two way distinction of English and other European languages. Moreover, the type of contrast is not between voiced and voiceless but among three manners of articulation called lax, tense, and aspirated. F0 after lax obstruents is low and F0 after tense or aspirated ones is high (Kim 1965, Han & Weitzman 1967, 1970, Hardcastle 1973, Kagaya 1974, Jun 1996) but reported observations on the order between the two manners, tense and aspirated, have not been consistent. However, all the experiments described in this thesis suggest that F0 is higher after aspirated sounds than after tense ones.

1.2 F0 and Automatic Speech Recognition

In modern automatic speech recognition, the exploitation of F0 is getting more frequent. Most F0 applications are, however, concentrated at the utterance level. To illustrate, a frequent use of F0 fluctuation is demarcation of syntactic boundaries using F0 information along with other prosodic cues such as pause and duration (Wightman & Ostendorf 1991, Nakai *et al.* 1996, Lee & Song 2000). Another application of utterance level F0 contours with spontaneous speech dialogue data is achieved by Taylor *et al.* (1996), King (1998), and Wright (2000), who show that word error rate can be reduced by automatically classifying utterance types using intonational characteristics combined with class specific language models. Their work is especially significant in that F0 is used directly to decrease word error rate in speech recognition. Both syntactic boundary demarcation and utterance type detection can be used for disambiguation of utterance meanings which is not possible with segmental information alone (Price *et al.* 1991, Hunt 1996, Hirose 1996, Jang *et al.* 1998).

Reports on successful exploitation of F0 for speech recognition at the segmental level are rare. Dumouchel & O'Shaughnessy (1993) and Dumouchel (1994) performed Hidden Markov Model (HMM) recognition experiments using segmental F0, together with duration and intensity. Diphones were used as basic recognition units and dynamic features for each of prosodic parameters were used for reducing speaker variability. To produce a final output probability for each observation sequence, the probability of segmental prosody was calculated independent of that of other acoustic features and the two probabilities were then multiplied. After tests with isolated word data, they report 4% recognition rate increase (from 48% to 52%). But the data set size and method of evaluation are unclear.

I have not found any report on Korean speech recognition exploiting segmental F0 influence.

1.3 Saliency of the Segmental F0 Effect in Korean

Though it is generally accepted that F0 of a vowel varies depending upon the type of the preceding consonant, there is a disagreement on whether the effect is strong and consistent enough to be practically exploited for the differentiation of consonant types.

Various reasons for skeptical conclusions are suggested. For illustration, the duration of the segmental effect is too short (Gandour 1974:346). The magnitude of the effect is not large enough to alter pitch perception and therefore it does not contribute to the functional roles of the prosodic attributes (Kompe 1997:97). Such F0 fluctuations are not intended by speakers and have no relevance for the perception of intonation (de Pijper 1983:14). Other imaginable reasons are that there can be a relatively large overlap between F0 values for different classes, and that the influence of other F0 fluctuation factors might mask the segmental effect.

However, most of the above critical remarks are based upon observations or experiments on European languages[1]. As for Korean, on the other hand, a cross-language phonetic experiment reports that the effect is considerably more salient than in English or French (Jun 1996). Her findings regard the magnitude of F0 and the duration of the effect. First, the magnitude difference between F0 values after strong consonants and after weak consonants[2] is significantly greater (by an average 50-80 Hz) in Korean than in English or French. Second, the effect lasts a lot longer in Korean (at least 100 msec) than in the other two languages (40-60 msec). Though she attempts to give an explanation for the effect itself in terms of perceptual saliency, she does not deal in depth with the reasons that make it more prominent in Korean than other languages. I will attempt to do so in the following chapter.

[1]Even in one of those languages (ie., English), Silverman (1987) argues against the criticism providing experimental evidence for the usefulness of segmental F0 effect.

[2]She classified voiced sounds and voiceless sounds into different classes for English and French, while aspirated, tense stops, and fricatives compose a single class distinguished from lax stops along with /m,l/. Thus, strictly speaking, the contrast she examined is not 'voiced vs. voiceless' but 'strong vs. weak'.

1.4 System Overview

Although it will be shown that segmental F0 values are useful for consonant manner identification, they supplement rather than replace other spectral cues. Thus, introduction of F0 will be conducted in a way that combines the two types of features together. One imaginable way of doing this is to explicitly separate the two parameter types using a two pass strategy. In the first stage, phones would be identified only by spectral parameters, as is done by a lot of the standard recognisers. Then the identity of each relevant phone could be refined by consideration of segmental F0. However, this method is intuitively unappealing even before taking into account the inefficiency of two-pass processing. It is very unlikely that in the human perception mechanism phones are roughly recognised in a first step and then the identities of only some specific classes of phones are readjusted based on other perceptual cues. Having to use thresholds to decide which initial rough hypotheses should proceed to the refinement stage is also clumsy.

Therefore, the method of implementation I adopted is to integrate F0 together with spectral parameters and have the recogniser use both to calculate the output probability of each phone. A difficulty in this method is to find a good unit of recognition, as segmental F0 is a cue extracted not from the consonant to be recognised but from a neighbouring phone. I suggest demisyllable units to cope with this problem. As demisyllables contain an onset consonant, if it exists, and a portion of nucleus in the same unit, all the features are extractable from the unit on its own. This will make it possible for phones, and subsequently words, to be determined without the awkward introduction of a separate F0 implementation step. Even before the introduction of F0, demisyllable recognition was found to be suitable for Korean speech recognition, as its performance was found to be better than a standard context dependent phone-unit system. This means that using demisyllable units has an independent motivation.

Hidden Markov Model training algorithms will be consistently used for most speech recognition experiments, as a number of speech recognition studies have proven that

they are quite powerful in probabilistic modelling.

In an effort to maximise the quality of speech recogniser, I modified a lexicon of Korean words. Possible variants are added to a baseform lexicon with a typical pronunciation, and infrequent ones are removed by probabilistic selection. The new lexicon with frequent variants is useful in all the recognition tests and the automatic method of pronunciation modelling seems to be potentially helpful for enhancing other speech recognisers or verifying existing phonological rules, not just in Korean.

1.5 Thesis Overview

In the immediately following chapter, I first provide the background of the research. It is useful to examine the physiological production mechanism of relevant obstruents in order to understand the reason for F0 variation depending upon manner of consonant articulation. This also helps in choosing the appropriate distinctive feature representation for Korean consonants. Then the reasons why Korean has a more prominent segmental F0 effect than European languages will be discussed. Further validation of earlier research is also described. A phone recognition experiment is performed to show that obstruent recognition needs to be improved. The types of errors in recognition of stops and affricates are also analysed to check out the possible scope for improvement through better recognition of stop and affricate manner.

In chapter 3, the construction of a baseline recognition system is described. As the baseline recogniser will be used as a benchmark to be compared with other recognisers with or without F0, modern conventional methods will be adopted for its training and test. Some of these methods will continue to be applied for other speech recognition experiments described throughout the thesis for feasible and legitimate comparisons. Details on the lexicon modification via pronunciation modelling are explained in this chapter as well.

The next two chapters cover the integration of the segmental F0 effect and speech recognition. In chapter 5, I describe how demisyllable units are created and implemented for Korean speech recognition. It is shown that although demisyllable recognition was initially motivated for the purpose of direct F0 implementation, it is valuable on its own since its quality turned out to be better than the baseline phone-based recogniser. Its performance will be regarded as another baseline to be compared with the model using F0. Chapter 6 covers the introduction of F0 into the recogniser. It will be shown how demisyllables are made to carry the characteristics of the segmental F0 effect. The results will be analysed and compared with those of competing systems.

Major findings will be summarised in chapter 7 followed by an indication of possible avenues toward further improvement using the segmental F0 effect.

Transcription of Korean phones and characters

To represent Korean pronunciation through the thesis, IPA symbols are basically used for readability. Notation for lax stops and affricates in each place are: p (bilabial), t (alveolar), k (velar), and c (palatal affricate). Corresponding tense stops are represented by adding a single quotation mark such as: p', t', k', and c', and aspirated sounds, by adding [^{h}] such as: p^{h}, t^{h}, k^{h}, and c^{h}. As some IPA symbols for other phones are not machine friendly, I have invented a few replacements. Details of phone transcription are shown in Appendix A on page 156.

When exemplifying Korean sentences, I use 7-bit Romanised characters instead of 8-bit Korean ones. The criterion and method of Romanisation is given in Appendix B on page 158.

CHAPTER 2

The Background, Problem, and Motivation

2.1 Introduction

The purpose of this chapter is to validate the research by describing the problem and providing supporting evidence for the usefulness of segmental F0 for automatic speech recognition.

First of all, however, a description of the segmental and prosodic structure of Korean is provided, as language particular characteristics of Korean are closely related to the saliency of the segmental F0 cues. The inventory of Korean consonants is illustrated in section 2.2, followed by a brief introduction to three consonantal manner types, in section 2.3. Related to this issue, physiological explanations for the different kinds of F0 perturbation associated with different obstruent manners will be provided in section 2.4.

Then, in the next two sections, I attempt an explanation for the reason why segmental F0 of vowels is an especially useful cue to the identity of preceding stops in Korean compared with European languages like English or French. First, I show, in section 2.5, how the simple syllable structure of Korean is helpful for associating vowel F0 with a particular onset consonant, followed by an account of the prosodic conciseness which

enables segmental F0 effects to remain reliable cues in spite of other prosodic factors influencing the fluctuation of vowel F0, in section 2.6.

One of the starting points for this study is the question of why there are few previous reports on the use of segmental F0 perturbation for automatic speech recognition, in spite of much empirical evidence for the existence of F0 as a cue to obstruent manner, provided through a number of phonetic experiments. One imaginable answer is that current speech recognition technology does not have any room for further improvement associated with the recognition of consonants. But I will show in section 2.7 that this is not the case.

In section 2.8, I analyse a list of Korean syllables, an online word dictionary, and transcription of a real speech data to show that obstruents are used in a large portion of Korean syllables and words so that the recognition of those obstruents is practically important for both human being and machine to capture the identity of words.

And finally, in section 2.9, cautionary remarks on the use of temporal characteristics, which have been claimed by many earlier phonetic studies as the most salient cue, are given for the purpose of emphasising the relative importance of the vowel F0 cue.

2.2 Typology of Korean Obstruents

The Korean consonantal chart in Table 2.1 is composed of 15 obstruents and four sonorants. Korean is a language which uses a relatively small number of continuant obstruents like fricatives in comparison to the number of non-continuants such as stops and affricates. Among 15 obstruents only /s/, /s'/ and /h/ are continuants and other 12 segments belong to the non-continuant category represented with the phonological distinctive feature [-continuant]. Among those three continuants, the glottal fricative /h/ does not have any homogeneous counterparts (ie., a phone contrastive only in voicing or manner) in the phoneme inventory and this makes it unnecessary to contrive yet another cue to help recognise that sound. No previous phonetic studies have reported F0 cues associated

	bilabial	alveolar	palatal	velar	glottal
Stop	p p' p^h	t t' t^h		k k' k^h	
Affricate			c c' c^h		
Fricative		s s'			h
Nasal	m	n		ŋ	
Liquid		l			

Table 2.1: Phoneme inventory of Korean obstruents.

with identification of /h/, either. Thus, the current study does not bother dealing with this sound in depth.

As for alveolar fricatives /s/ and /s'/, it has been frequently assumed that their distinction is in line with the contrast of stops and affricates in spite of the fact that only a two way manner classification is available rather than the typical three way contrast of stops and affricates. Namely, /s/ is regarded as a lax segment while /s'/ is classified as tense. But a few acoustic experiments suggest that this does not seem to be an appropriate distinction (Kagaya 1974, Park 1999). These studies will be discussed in the next section dealing with physiological aspects, as their arguments are based on laryngeal gestures.

The allophones of stops and affricates and examples of contrastive usage in words are presented in Table 2.2. In syllable final position followed by a pause, only unreleased variants may occur regardless of manner difference, so all the bilabial stops are realised as [$p^\urcorner$], alveolar stops as [$t^\urcorner$], and velar stops as [$k^\urcorner$]. In this context, affricates (and fricatives) are neutralised as [$t^\urcorner$]. The F0 cues studied in this thesis are located in vowels immediately following obstruent consonants. As the unreleased variants are never directly followed by vowels or syllabic consonants, no such F0 cues can occur. On the

Type	Phoneme	Allophones	Example
Lax	p	p b p' p^h β w $p^\neg$	pul 'fire'
	t	t d t' t^h $t^\neg$	tal 'moon'
	k	k g k' k^h ɣ $k^\neg$	keda 'to tidy'
	c	c z c' c^h $t^\neg$	cim 'burden'
Tense	p'	p' $p^\neg$	p'ul 'horn'
	t'	t' $t^\neg$	t'al 'daughter'
	k'	k' $k^\neg$	k'eda 'to break'
	c'	c' $t^\neg$	c'im 'steaming'
Aspirated	p^h	p^h $p^\neg$	p^hul 'grass'
	t^h	t^h $t^\neg$	t^hal 'mask'
	k^h	k^h $k^\neg$	k^heta 'to dig'
	c^h	c^h $t^\neg$	c^him 'needle'

Table 2.2: Phoneme, allophones, and examples of stops and affricates. Diacritics used for narrow transcription are: (') for tenseness, (^{h}) for aspiration, ($^\neg$) for unreleasing. Source of allophones: Huh (1991).

other hand, since the distinctions used by F0 intervocalically are neutralised anyway, the cues are not needed.

When surrounded by vowels or other voiced sounds, lax stops can be realised as voiced counterparts [b, d, g], voiced fricatives [β, ɣ], or even an approximant [w] in the case of bilabials. It should be noted that aspirated sounds and tense sounds are not realised as voiced variants in any case. As only lax sounds can be phonetically voiced, and F0 of vowels is known to be useful for voicing distinction[1], the segmental F0 effect seems to be useful for manner differentiation of stops in intervocalic position, too. That is, if an intervocalic obstruent is detected and if it is voiced, its manner of articulation can be easily decided to be 'lax'.

The domain of the intervocalic lax stop voicing assimilation appears to be further restricted. When a lax obstruent is located at the edge of a prosodic unit which is larger than "prosodic word", it does not undergo the assimilation and remains voiceless. In such

[1] Voiced sounds lower the F0 of the following vowel and voiceless sounds raise it.

cases, phonetic voicing itself cannot act as a cue to obstruent manner, but the F0 of the following vowel still can. Further consideration of this issue will be discussed later in this chapter and across other chapters.

The three way distinction of obstruents appears to be closely related to the discriminative role of segmental F0 perturbation. Korean speakers must distinguish the three manners of articulation phonetically as well as phonologically, whereas, in comparison, English speakers settle for two-way voicing distinction. As Korean speakers, in the process of obstruent production, cannot make use of the acoustically more feasible two-way voicing distinction, they are forced to rely on other three-way cues like segmental F0 and/or temporal characteristics.

2.3 Representation of Three-way Distinction

Through cross language studies, including Korean, of stops Lisker & Abramson (1964) argue that the time difference between the release of articulatory occlusion in the vocal tract and the laryngeal gesture of voicing, so called Voice Onset Time (VOT) is the primary cue to distinguishing manner contrasts of stop categories. They state that other conventional features like voicing, aspiration, and force of articulation are likely to be predictable from that primary cue.

Against Lisker & Abramson's idea on feature priority, Kim (1965) argues that force of articulation, or *tensity* is also a necessary criterion for classifying stop consonants, especially in Korean. Based on a series of phonetic experiments, he showed that this feature is the key to separate out the lax stop class from the other two. Thus, he includes aspirated stops in the category *tense* as well as the class composed of the sounds in the second row in Table 2.2 which I call *tense stops* through this thesis.

I believe this initial insight is closely related to the fact that the lax stops give rise to lower F0 in the following vowel than the other categories of stops. But a binary feature

	Lax	Tense	Aspirated
Spread Glottis	+	−	+
Constricted Glottis	−	+	−
Stiff Vocal-cords	−	+	+
Slack Vocal-cords	−	−	−

Table 2.3: Laryngeal features of stops and affricates in word-initial position. From Halle & Stevens (1971).

of tensity alone is certainly not sufficient to represent the three way distinction of stop categories.

The three way distinction is effectively represented in terms of features which reflect various aspects of laryngeal gestures, suggested by Halle & Stevens (1971). They are summarised in Table 2.3. Vocal cord slackness is highly correlated with the voiced/voiceless distinction, and all the classes are represented with [-slack vocal cord], which properly describes the fact that all Korean stops are voiceless at word initial position (compared with English /b/ with [+slack vocal cords]). While slack vocal cords invoke vocal cord vibration, stiff vocal cords prevent it. Only the lax category has [-stiff vocal cords] and this accounts for the fact that only this category of stops undergoes voicing assimilation, producing various voiced allophones as illustrated in Table 2.2.

In agreement with Kim's findings on glottal features, Halle & Stevens attribute aspiration to the degree of glottal opening, imputing [+spread glottis] to aspirated sounds. In fact, another difference between the Korean lax stops and English voiceless stops is that the former have slightly more aspiration than the latter. This is captured in the above feature specification where lax stops are given [+spread glottis]. Of course, the extent of spreading is distinctively greater for aspirated stops and this is not describable with a binary feature. Instead, the aspirated stops can be distinguished in terms of the feature vocal cord stiffness. Finally, tense stops are unique as to the feature [+constricted

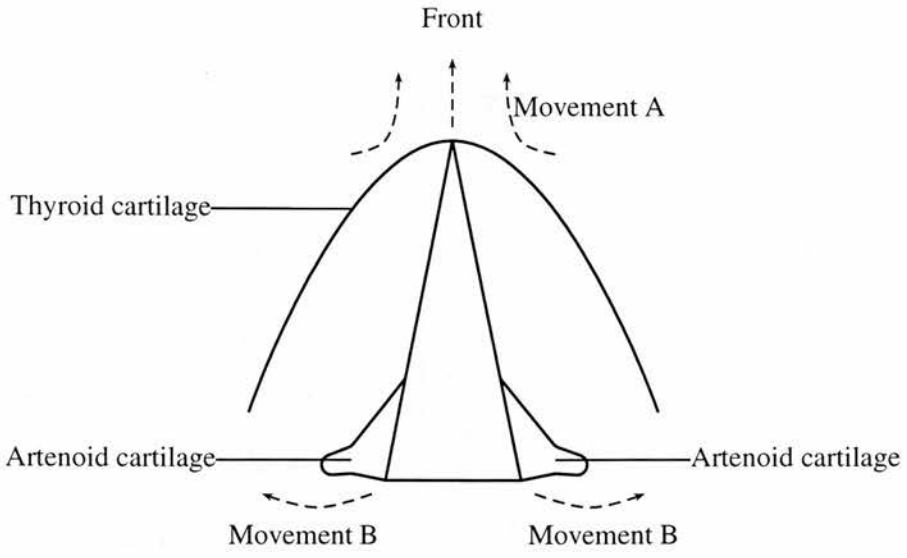


Figure 2.1: Schematic view of the larynx from above. Movement A is associated with pitch feature and movement B is principally related with glottal stricture. Source: Ladefoged (1973:74), Laver (1994:185).

glottis]. I will follow this feature specification of Korean stops not only for their descriptive convenience but because of the physiological plausibility of accounting for the cause of different segmental F0 perturbation for each obstruent class, of which the discussion immediately follows.

2.4 Laryngeal Gestures and Segmental F0

Ladefoged (1973) states that the main physiological condition of pitch raising is associated with the stretching of the glottis in the anterior-posterior dimension (the movement A in Figure 2.1). On the other hand, the pitch fluctuation caused by different consonantal types needs a different physiological explanation, since it is not so much the anterior stretching of glottis as the various conditions involving the arytenoid cartilage (the movement B in Figure 2.1) that determines the type of obstruents pronounced. These include intrinsic muscle movement, vertical disposition of the larynx, size of the glottal aperture, transglottal air pressure or a mixture of one or more such factors.

Aspirated

An appealing explanation for the high F0 after an aspirated sound is in terms of aerodynamics (Ohala 1978). As already mentioned, when aspirated sounds are articulated the arytenoid cartilages are wide spread apart (ie., [+spread glottis]) during the oral occlusion due to the movement B in Figure 2.1. At the moment when the occlusion is released, the oral pressure decreases rapidly causing the rate of airflow through glottis to increase. Thus increased velocity of the air causes the Bernoulli effect by which the vocal cords will be drawn together and pushed apart (ie., vibrate) at a high speed.

Tense

The cause of high F0 after tense stops is quite different from that of aspirated stops. In terms of the aerodynamic account on its own, Korean tense sound articulation would lower the pitch instead of raising it because the glottis remains tightly closed (ie., [+constricted glottis]) by the inward movement of arytenoid cartilages (the opposite direction of movement B in Figure 2.1). This constricted phase of the glottis during the tense stop articulation has been physically verified by Kagaya (1971) with pictures taken through a fiberscope inserted into the speaker's nostril. Thus, the rate of airflow immediately after release decreases a lot compared with the articulation of aspirated sounds or even of lax sounds. But as Ladefoged (1973:74) indicates, there is not just one physiological correlate of pitch variation. It appears to be muscular tension that causes F0 raise in the articulation of tense stops. Stiffening of the vocal cords is known to affect the frequency of vibration regardless of the size of the glottal aperture (Halle & Stevens 1971:202). The tightened glottis during the occlusion inhibits the vibration of vocal cords. But when the oral closure is released, the stiffness is maintained for a little while and facilitates the vocal cords' vibration at high frequency. An observation of muscular activities through electromyographical investigation supports the increased stiffness of tense sound articulation:

> Both muscles [the vocalis and the lateral cricoarytenoid], VOC[the vocalis] in particular, showed a marked increase in activity before the stop release in type I [the tense type], which presumably resulted in an increase in inner tension of the vocal folds. (Hirose *et al.* 1974, cited in Kagaya 1974)

Lax

As for lax stop articulation in the word initial position, the widening of the glottis does take place (ie., [+spread glottis]) as physically observed by Kim (1970), but the degree of abduction is not as much as in the case of aspirated sound articulation and there is no particular muscular tension (ie., [-stiff vocal cords]) or glottal constriction (ie., [-constricted glottis]). In other words, there is no element which raises F0 in the lax stop articulation.

Though it is relatively easy to describe why tense stops and aspirated stops lead to higher pitch than lax stops, it is not obvious why vowels after aspirated stops have higher pitch than after tense stops since they have different mechanisms of pitch raising. One clue derivable from phonetic studies is the size of the subglottal air pressure which causes the high rate of transglottal airflow. Although it seems to be the case that both tense and aspirated sounds have have considerably higher subglottal air pressure than lax sounds (Chomsky & Halle 1968), pressure for aspirated sounds seems to be relatively higher than for tense sounds. The experiments conducted by Lee & Smith (1972), and Kagaya (1974:175) confirm this. As a consequence, it can be inferred that the segmental F0 effect in Korean is affected to a greater degree by the high air pressure factor than by the muscular tension factor.

Affricate

The above consideration of the interaction of F0 and obstruent articulation is mostly based on the literature on stop sounds, but the articulation of affricates can be explained in more or less the same manner as for stops. As supporting evidence, Kagaya (1974)

shows that the glottal width of aspirated affricates at the point of release is larger than for the other two classes, which invokes rapid air flow and subsequently higher vowel pitch. The tension seems to be highest when a tense affricate is produced due to the constricted glottis. Therefore, the factors which increase the F0 are common to both stops and affricates. It will be shown by experiment, on page 87, that the segmental F0 effect is at least as salient in affricates as in stops.

Fricative

The alveolar fricatives /s/ and /s'/ are unusual in that they show only two way contrast rather than the typical three way contrast of other non-continuant sounds. There is no controversy in classifying /s'/ into the 'tense' category as its acoustic properties have much in common with tense stops or affricates. As for /s/, it has been traditionally grouped with lax stops, probably because of the Korean orthographic convention for /s/, which is analogous to the way lax stops and lax affricates are represented. But Kagaya (1971) argues that /s/ is closer to aspirated stops rather than lax stops in acoustic quality. His fiberscopic investigation shows that the change of glottal width of that sound during the course of articulation is similar to aspirated stops. This implies that there is a substantial amount of aspiration while /s/ is pronounced, which is also confirmed in a more recent study (Park 1999). Another piece of evidence on /s/ as an aspirated sound is that it, unlike lax stops, does not undergo the voicing assimilation rule in intervocalic position. As has been mentioned before, this behaviour patterns with aspirated stops rather than lax stops (however, see Iverson 1983). The investigation of vowel F0 behaviour after fricative sounds described in section 6.3.7 on page 133 is also supporting evidence. Namely, the F0 value of a vowel following /s/ is significantly higher than following /s'/, as is the case for tense stops versus aspirated stops. Therefore, I will regard the Korean fricative /s/ as an aspirated fricative rather than a lax fricative. However, it is still premature to conclude that the segmental F0 effect can be exploited for the recognition of these fricatives. There is an independent source of difficulty. At normal speech rates, vowels are very often deleted or devoiced when preceded by those sibilants, as described in section 2.5.

Thus, the extraction of F0 itself, and hence its exploitation is impossible. This implies that Korean speakers cannot rely much on F0 in the distinction of /s/ and /s'/.

2.5 Syllable Structure

The simplicity of Korean syllable structure also supports the consonant-vowel connection in articulation. Adjacency of consonant and vowel needs to be maintained for the vowel to retain the consonantal effect in its F0. When more than one consonant is positioned before a vowel, it becomes more difficult for a consonant, which is not immediately adjacent to vowel, to coarticulate with the vowel. For example, in an English syllable composed of three onset consonants before the nucleus, the acoustic information of the first consonant is not transmitted readily onto the nucleus position across other consonants.

On the other hand, the simple segmental structure of Korean syllables makes it easy for a vowel to contain postconsonantal F0 information. In Korean, phonemically only one consonant is allowed at the both onset and coda position in a syllable restricting possible syllable types to one of [2]:

CV, CVC, VC, V

Among the above four syllable types, the number of syllable tokens of the first two types (ie., CV and CVC) is far greater than that of the other types. The ratio is found to be 95% vs. 5% (Huh 1991:229)[3]. Therefore, it is obvious that Korean syllables sound more natural when pronounced with their onset position filled.

The single onset C slot allowed implies less complex segmental realisation at the surface phonetic level in comparison to other languages with consonant clusters allowed in their syllable onset position. As a consequence, the occurrence of consonant clusters,

[2]The orthography of words with two successive consonants in the coda position, such as *alm* 'knowledge' or *kaps* 'price', does not imply the phonemically two consonants. As only one consonant of the those two is realised in pronunciation ([am], [kap], respectively.), the complex coda representation in orthography is generally regarded as reflection of the morphophonemic structure (Kim 1965, 343, footnote 22).

[3]The ratio is obtained from all theoretically possible syllable tokens of spoken Korean.

Context in Syllable	Number	Percentage (%)	Example
C _ sonorant	163	45.40 (2.09)	p^h u m → p^h ϕ m
fricative _ $	135	37.60 (1.73)	s i → š ϕ
stop/affricate _ $	61	16.99 (0.78)	k i → k ϕ
Total	359	100 (4.59)	

Table 2.4: Comparison of vowel deletion contexts: percentage refers to the ratio to total number of affected syllable, and in the parentheses, the ratio to total number of syllables. For example, 45.40% of vowel deletion occurs when the vowel is followed by a sonorant. '$' is a syllable boundary marker.

which makes it harder to relate consonantal identity with acoustic features at the following vowel, is substantially constrained. In other words, cohesion between onset and nucleus is stronger in Korean than in the languages like English or German which allow association of up to three C slots to the syllable onset position (eg. [str], [spl]).

In spite of these strict phonological syllable structure constraints which are expected to subsequently prevent, to a certain degree, the phonetic occurrence of consonant clusters, the structure is not entirely preserved and consonant clusters do appear at the phonetic level, as a result of postlexical phonological rules like vowel deletion. The extent to which the deletion process happens is dependent upon how an utterance is spoken. The more naturally and the faster the speech is produced, the more likely vowel deletion is to take place. If this phonetic process happens very often, it can be a threat to the exploitation of segmental F0 for consonant detection because it is available only when the consonant is in the vicinity of a vocalic segment.

To find out how often the original syllable structure is transformed by vowel deletion I examined a portion of a continuous speech database, which is described in detail on page 50. Phonetic hand labels for 433 utterance tokens are examined (See page 52 for detailed description of labelling).

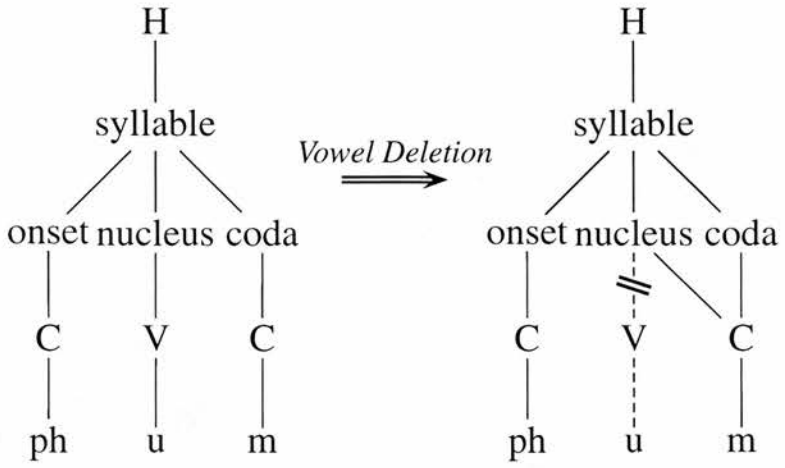


Figure 2.2: Vowel deletion and tonal association with syllabic consonant.

The dataset contains a total of 7816 syllables of which 359 (4.59 %) syllables underwent vowel deletion. Table 2.4 shows how deletion applied depending upon the segmental structure of the syllable.

Of all the original syllables that are affected by vowel deletion, 45.40% include a syllabic consonant in coda position. Strictly speaking, some of these vowels are not so much deleted as totally nasalised, but when the vowel period is found too short to be separately marked, the vowel is regarded as deleted.

When the syllable structure is CVC with a sonorant in the coda position, vowel deletion does not appear to critically undermine the analysis of segmental F0 since the tonal characteristics of the nucleus might be transmitted to the next sonorant instead of being removed along with the segment. The non-linear hierarchical structure of *Autosegmental Phonology* (Goldsmith 1976) provides the theoretical background for such an explanation. As the tonal tier and the syllable tier have independent status and the two tiers are related to each other only in terms of association, it is the syllable node itself rather than each segment which carries the tonal property. Thus, the deletion of a vowel from

a syllable can be interpreted as dissociation between the nucleus node and the V node. Subsequently, the empty nucleus node is reassociated with the C node in the coda position to satisfy the well-formedness of the syllable structure. In other words, when the coda segment becomes an alternative syllable nucleus, the tonal property is mainly realised through that segment without losing its characteristics altogether. This process can be schematised in Figure 2.2.

This explanation is supported by empirical investigation. An analysis described in section 6.3.5 on page 129 confirms that the tonal behaviour of syllabic consonants following vowel deletion sites also reflects the type of previous obstruent in the same way as vowels. Thus, vowel deletion before sonorants does not seriously obstruct the use of the segmental F0 cue.

Of the other 196 vowel-deleted syllables without any alternative tone bearing segment, 135 cases have fricative sounds ([s] or [s']) in their onset position. They are not obstacles to the current research which deals with only non-continuant obstruents. Thus, only a small number (61 cases, that is, 0.78% of all syllables or 1.82% of the syllables with a stop or an affricate at onset position) of vowel-deletion cases can be problematic by making it impossible to extract the vowel F0.

A possible explanation for the lack of deletions after a stop or affricate is directly related with the role of vowel in identifying the previous consonant. Absence of the vowel will undermine the discrimination of the preceding consonant and increase ambiguity. To prevent this, vowel deletion is suppressed. On the other hand, the deletion process is less restricted when preceded by fricatives /s/ and /s'/, because the vowel does not contribute much to the differentiation of those consonants. This tendency is another piece of supporting evidence for the association of vowel F0 and non-continuant classification.

2.6 Prosodic Structure

Apart from segmental F0, there are a variety of factors which influence the F0 of vowels. In order to utilise segmental F0 for phone identification or further speech recognition, it is crucial to separate out its effect from such factors. Otherwise, the segmental contribution to F0 may become hard to detect because it can be masked by other factors.

As was already mentioned in the Introduction, Jun (1996) has shown that segmental F0 is more salient in Korean than European languages. I believe one of the reasons for this is the relatively simple and phonetically predictable structure of Korean prosody in general. In this section, I describe how segmental F0 factor is related or unrelated to other F0 varying factors.

I will begin with the brief description of Korean prosodic structure in general as its understanding is essential to examine the interface of factors.

2.6.1 Prosodic hierarchy

A generally accepted agreement on prosodic structure is that there is a hierarchy among the prosodic units of a language such that a unit is exhaustively and exclusively composed of one or more units in the immediately lower level. Nespor & Vogel (1986:7) states that this process of grouping prosodic categories is realised according to a language universal rule that they call *Prosodic Constituent Construction*, while specific well-formedness conditions apply to each language such as *Strict Layer Hypothesis* (SLH) (Selkirk 1984:26) of English.

Those notions are also useful in describing Korean prosodic structure, which I assume is hierarchically constituted as:

(2.1) Utterance
Intonational Phrase
Accentual Phrase
Prosodic Word
Syllable

For example, an *Utterance* consists of one or more *Intonational Phrases* (IP) which consists of one or more *Accentual Phrases* (AP), and so forth. Both in Selkirk (1984) and Nespor & Vogel (1986), the constituent level between IP and *Prosodic Word* is not AP but *Phonological Phrase*. Jun (1993) replaces it with AP on the ground that this category is better determined based on intonational characteristics rather than on the concept of syntactic boundaries.

Among the above categories, IP and AP are defined mainly on the basis of their intonational structure. Thus, it is possible that the F0 contours of those prosodic units are closely related to segmental F0 effect in one way or other, so more detailed description for AP and IP is necessary.

Accentual Phrase

The term *accentual phrase* is introduced by Beckman & Pierrehumbert (1986) in their comparative analysis of Japanese and English intonation structure. They describe AP as "the lowest level of phrasing that is well defined by the intonation pattern"(page 261). Jun (1993, 1998) extends this notion to Korean prosody and formalises it on the basis of various phonetic experiments. She justifies the usefulness of AP showing that some phonological rules of Korean, such as *obstruent nasalisation* or *post obstruent tensing*, are better described using AP as their application domain.

The AP internal tonal patterns are important with regard to the segmental F0 perturbation. Jun states that the underlying tonal pattern of AP in Seoul Korean is either LHLH or HHLH depending on the segmental structure of the AP. These two tonal patterns realise as

Number of syllables	Tonal patterns
1-2	XH
3	XH, XLH, XLH, XHH
4 or more	XHLH
	(where X is L or H)

Table 2.5: Types of phonetically realised AP from Jun (1998). The first tone (shown as X) is determined depending upon the type of onset consonant in the first syllable of AP (ie., segmental F0 effect).

a variety of other shapes depending on the number of syllables in the AP, as in Table 2.5. An AP is composed of maximum of four tonal slots, even though the number of syllables in one AP can be larger than four.

The most decisive tonal element of AP is the last H tone located at the right edge of the phrase. Note that all the variants of AP tonal pattern have this phrasal tone.

Below is a typical example of Korean accentual phrasing (for more examples of accentual phrasing see Jun (1993:42ff)):

(2.2) *[na_ncun]*$_{AP}$ *[neo_leul]*$_{AP}$ *[co_a_hae]*$_{AP}$
I-TOP you-ACC like-END
'I like you'

The APs in the above example have default tonal patterns LH, LH, LLH, respectively. In fast speech, two or more APs can be collapsed together. For example, the first and second APs, or the second and third APs, in the above example can constitute a single AP.

It is the segmental F0 effect that determines the shape of the initial tone, which is represented as 'X' in Table 2.5. Namely, when an AP starts with a tense or aspirated obstruent (ie., [+stiff vocal cords], in Figure 2.3), the AP initial tone is realised as H, otherwise L. Thus, the default tonal patterns of the first and third APs in Example 2.3 are both HH rather than LH as they begin with aspirated and tense sounds respectively.

(2.3) *[cha_neun]$_{AP}$* *[u_li_po_ta]$_{AP}$* *[ppal_la]$_{AP}$*
car-TOP than_us fast-END
'The car is faster than us'

There is also a tendency for the second syllable of an AP, which is composed of three or more syllables, to also have an H tone (eg., the second tone of XHLH in Table 2.5). But this tone is rather flexible both in its quality and position. When an AP is composed of fewer than three syllables, this medial H tone is undershot and does not appear phonetically. In three syllable APs as well, the undershooting appears to be common. The appearances of the AP medial tone (eg., the second tone of XHH) in three syllable APs can be attributed to other causes. First, when the onset position of the second syllable is tense or aspirated, that syllable bears H tone irrespective of the first syllable tonal pattern. Thus, either LHH or HHH can be realised depending on the onset type of the second syllable. When the onset of the second syllable is empty or occupied by a lax consonant, it is the first syllable tonal shape which determines the tone of the second syllable. If the first syllable has H tone due to the syllable onset consonant type its effect seems to persist in the second syllable which would otherwise bear L tone. This H tone spreading across the syllable boundary has been confirmed through a phonetic experiment by Lee (1998).

In APs with four or more syllables, the intermediate H tone is more likely to appear. But even in those syllables it may be suppressed in fast speech (Jun 1998:6). In an investigation of continuous speech data (for detailed description of data see page 50), I confirmed that in many long APs the intermediate H tones are suppressed or considerably weakened in magnitude and consequently intermediate H tones do not appear. Therefore, it seems that the only genuine property which enables phonetic determination of an AP is its phrasal tone located at the right edge.

The most important observation about AP, for purposes of the current research, is the consistency of the left edge tone which is represented as 'X' in Figure 2.5. As has been described above, the phrasal tone and possible medial H tones are not directly relevant to the left edge position. As far as the AP formulation is concerned, the only factor that

characterises the tonal shape of the first syllable position is the segmental F0 effect. This makes it possible for F0 of AP initial position to be used as a cue for differentiating the consonant type at the syllable onset position. This characteristic will be fully taken advantage of in the ASR implementation in chapter 6.

One related question is the constituency of AP. How many prosodic words usually constitute an AP? Though there are, as stated by Jun (1993), various factors of accentual phrasing such as speech rate, focus, or syntactic constraints, I found that a vast majority of APs in real data (94% in a subset of the database) are composed of only one prosodic word. This observation will be effectively taken advantage of in the later speech recognition experiments using segmental F0 effect.

Intonational Phrase

The Intonational Phrase (IP) has been widely agreed to be one of the language universal prosodic units. Though it is frequently defined in relation to syntax (Nespor & Vogel 1986), semantics (Selkirk 1984), or their interaction (Pierrehumbert 1980), the main interest of the current research lies in its phonetic properties with special regard to whether they interact with the segmental F0 effect. Various types of local F0 perturbation are related to the internal structure of IP, but an IP is most often characterised by the patterns of boundary tone usually observed at the end of it, which play a crucial part in determining the meaning of utterances.

In Korean, IP is a level higher than AP in the prosodic hierarchy. This means that one IP can be broken down into one or more APs. The right boundary of the last AP is always aligned with the right boundary of the IP. This alignment causes inevitable superposition of the AP phrasal tone and the IP boundary tone. When overlapping of tonal patterns of different level units occurs, the one associated with the unit at the higher level always wins. That is, the IP boundary tone preempts the AP phrasal tone. It should, again, be noted that those IP boundary tones do not directly interact with the segmental F0 effect which is supposed to be at the left edge of AP.

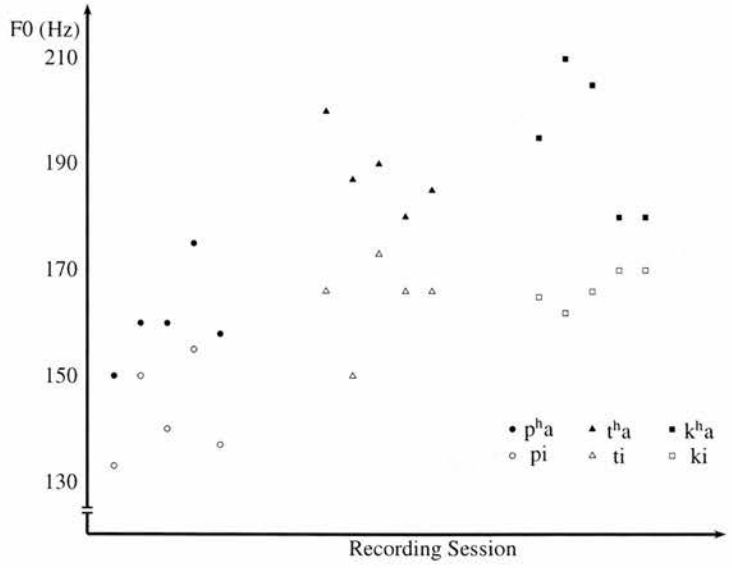


Figure 2.3: Comparison of interaction between IF0 effect and segmental F0 effect of a single speaker. The F0 values are taken from Han & Weitzman (1967:16). They state that the measurements were taken from "onset of each vowel" but the exact range of measurement is not clear. Each point in the graph represents the F0 value of a single utterance token. Each pair of vertically aligned tokens were uttered in the same recording session. It is shown that, in each session, the F0 value of the syllable composed of [aspirated stop + low vowel] is always greater than that of the syllable composed of [lax stop + high vowel], which means that the IF0 effect never completely preempts the segmental F0 effect.

We have seen that the most important characteristics which define main prosodic units like AP and IP are not located where the segmental F0 effect is most valuable. However, there are other prosodic factors that influence the intonational contour of an utterance, a few of which I will examine independently.

2.6.2 *Segmental level: intrinsic vowel F0*

Many studies have found that vowels have an intrinsic difference in F0 magnitude which covaries with the height of the tongue when they are articulated. This effect is usually called *intrinsic F0* (IF0) or *intrinsic pitch* of vowel. That is, high vowels like [i], [u] have higher F0 while low vowels like [a], [æ] have lower F0. As this property has been testified

over many languages, it is generally accepted as one of the universal characteristics of vowel quality (Whalen & Levitt 1995, Silverman 1987).

Korean is not an exception. Han & Weitzman (1967) measured F0 of vowels, produced by two people, in isolated nonsense syllables beginning with various stop sounds. The average difference between high and low vowels is approximately 20 Hz for a male subject or 11 Hz for a female subject[4]. This vowel quality can be a problem for the use of consonantal difference on F0, since the two factors may be mixed up together cancelling each other's effect. For instance, the difference of F0 between a high vowel preceded by a lax consonant (eg., [pi]) and a low vowel preceded by an aspirated consonant (eg., [p^{h}a]) may be obliterated if the IF0 effect is high enough. However, this does not seem to happen very often. The measured F0 values for each token do not reveal any significant overlap between the type 'aspirated stop + low vowel' and the type 'lax stop + high vowel'. Figure 2.3 is given for visual inspection of this phenomenon. The measurements of male voice were used for this illustration as they show stronger IF0 effects in the magnitude as mentioned above. It is shown that variation does exist across sessions but it never happens that the IF0 effect preempts the segmental F0 effect.

Though the existence of IF0 is confirmed by Han & Weitzman's data, it is still premature to draw any conclusion on the size of the effect as only one subject for each gender is involved in their experiments on IF0. I conducted more a comprehensive investigation on this issue using a larger amount of data, whose result will be given in chapter 4 and 6.

2.6.3 Word-internal level: stress

The next level to consider is the word-internal domain. If a language has lexical level stress or tone which makes a word phonologically contrastive, that prosodic factor can obstruct the segmental F0 effect as F0 is found to be one of the primary acoustic correlates of stress (Lea 1977). The effect may still be maintained but its magnitude becomes either

[4]Different results between the two speakers cannot be generalised as gender difference, as other experiments, whether for Korean or other languages, do not show the systematic IF0 difference between genders.

further amplified or attenuated based on whether the relevant syllable is stressed or not, or the duration of the effect may be restricted to a very brief period.

There is general agreement that there is no such phonemically contrastive stress or tone in Seoul Korean. But it does not necessary mean that there does not exist any relative prominence between syllables. Though irrelevant to the identity of words, it has often been pointed out that there are lexical accents in the Korean words.

Huh (1991:246) states that "there is a general tendency in Korean that the first or often the second syllable of a prosodic word is pronounced strongly, though the extent is not striking" (my translation). But it is clear that he, by the expression "strongly", specifically refers to intensity of the syllable rather than pitch. In the next page of his article, he mentions the pitch separately to note that no consistent behaviour is found in the pitch of a word.

H. B. Lee (1973:11-14)(cited in Lee 1990) and H. Y. Lee (1990) present the specific stress rules for Korean words. Their point, relevant to the current research, can be summarised as: if the first syllable is heavy it is stressed. By "heavy syllable", they mean either a syllable with coda position filled (ie., (C)VC) or with long vowel (ie., (C)V:). Especially, H. B. Lee (1973) adds that the duration of the vowel is the most important component of Korean accent and that pitch is not essential. (cited in Lee (1990:42)).

As a matter of fact, some of the above observations on Korean word stress might be attributed to the segmental F0 effect. For example, some data used to show the existence of first syllable stress contain the tense or aspirated consonants at the onset position of the first syllable.

Based on the above discussion, it is expected that non-distinctive lexical stress of Korean, if it exists, does not affect the segmental F0 very much.

2.6.4 *Semantic aspect: focus*

In an utterance, some constituents are more prominent than others, which can be explained in terms of the notion *focus*. In many cases, focus seems to be imposed on individual word (narrow focus), but Ladd (1996) shows that the higher level constituents like phrase can also be focused (broad focus). The distribution of accents in the focused constituents is the major concern with respect to current research since it might interfere with the segmental F0 effect. Especially, it is known that focusing influences the accentual phrasing of Korean utterances in such a way that every focused word initiates a new AP (Jun 1993, Oh 1999), which is illustrated as follows:

(2.4) a. *[uli komo_ka]$_{AP}$ [apha]$_{AP}$*
our aunt-TOP sick-END
'My aunt is sick'

b. *[uli]$_{AP}$ [**komo_ka** apha]$_{AP}$*
our aunt-TOP sick-END
'My *aunt* (not my uncle) is sick'

Example (2.4a) is a normal sentence without any particular focus put on the word [komo] 'aunt'. But (b) is a reply to a question like 'Is it true that your uncle is sick?'. In this case, focus is put on the word [komo] as an intention to correct the wrong idea of the dialogue partner. A normal AP phrasing without focus would be like that shown in Example 2.4(a)[5], combining the first two prosodic words into a single AP. But when the second word [komo] is focused it aligns with the left boundary of a new AP (Example 2.4(b)). But that is the very position where segmental F0 should be distinguishing stop manner. Therefore, if F0 is one of the main acoustic correlates of focus, and if the accent of a focused word is put on the initial syllable, it could interact with segmental effect making it hard to disentangle one from the other.

In lexical-stress languages like English, focus will be realised mainly in association with the lexically-stressed syllable in a word. As already discussed in the previous subsection,

[5]But separate phrasing of two words, such as [uli]$_{AP}$ [komo]$_{AP}$, is also possible.

Korean does not have distinct lexical stress. So we must consider how focus is realised on a word.

Recently, Oh (1999) performed an experiment related to this problem. She obtained 160 Korean sentence tokens (8 sentences x 4 speakers x 5 repetitions), half of which were spoken with focus on target words while the other tokens were spoken as a normal sentence. She compared the F0 and duration of each syllable in target words. Her results show that "F0 values for focused words is higher than F0 values for neutral words but the difference between them is not [statistically] significant when both are accentual phrase-initial" (Page 1518). This implies that focus does not crucially change the magnitude of F0 in AP initial position. On the other hand, she shows that it is vowel duration which is mainly affected by focus. The initial syllables of focused words were significantly longer than those of corresponding non-focused words. This is obviously an encouraging indication for the current research as the begining of AP is the position where the segmental F0 effect will be most effectively used.

2.6.5 Declination

Finally, at the utterance level, the tendency for F0 to gradually decline during the course of an utterance, often described as 'declination' is a troublesome factor, as it affects the local magnitude of F0 as a function of time. If the amount of the declination can be represented by a linear slope, rough normalisation of the F0 points can be achieved making it possible to compare segmental F0 through the utterance. More discussion on this issue will be presented in chapter 6.

2.6.6 Summary

It appears that the segmental F0 effect in Korean utterances is still maintained even after the major prosodic structure is established. There are two main reasons. First, Korean prosodic structure is found to be relatively simple. There is no phonologically contrastive tone or stress, so the degree of complication with underlying intonation is also relatively

small. Second, the location where segmental F0 perturbation is well observed and the points where other F0 influencing factors are most prominent do not appear to coincide very often. The major prosodic units defined by intonation, such as AP and IP, have their crucial characteristics on the right edge whereas the segmental F0 effect is found to be most effective at the left edge of prosodic units.

The F0 factors discussed in this section are by no means exhaustive. And some factors are not satisfactorily investigated with various other environments fully taken into account. Therefore, more empirical verification will be always helpful and some will be conducted in the later chapters.

2.7 Low Recognition Accuracy of Stops and Affricates

In this section, I will show that the automatic recognition of relevant stop and affricate sounds needs to be improved, based upon a result of a phone recognition experiment.

I confine the consideration in this section to an experiment of phone unit recognition. Phone recognition performance is in close correlation with word recognition performance. Although it is the case that well organised word sequence grammar models, based on a high level artificial grammar or a statistically calculated probability, play a very important role in the modern ASR systems, acoustic probability calculated directly from the signal must provide the fundamental source of information in decoding a given utterance signal (Koreman *et al.* 1997:85).

The phone recogniser used in the following experiment is constructed in roughly the same way as a continuous word recogniser is created, simulating each phone token as a word token. As the method by which a continuous word recogniser is constructed is described in detail in chapter 3, only a brief summary is given here. 37 (36 phones + silence), three state, left to right, continuous density Hidden Markov Models are generated to represent each phone using HTK, (HMM Tool Kit, Young *et al.* 1996) which adopts the *Baum-Welch* training algorithm for modelling and *viterbi* algorithm (Forney 1973) for

Phone Class			No-gram	Unigram	Bigram
Vowel			50.67	53.32	61.98
Consonant			59.61	59.31	64.95
	Sonorant		62.57	60.25	66.22
	Obstruent		57.36	58.59	63.98
		stop	**49.61**	**50.56**	**56.12**
		affricate	**57.68**	**58.42**	**65.85**
		fricative	73.97	76.13	79.74
Overall average			56.97	58.81	62.70

Table 2.6: Summary of phone recognition accuracy(%) for natural classes. Accuracy is calculated through the formula: $((Correct - Insertion)/Number) \times 100$.

recognition. 3000-word speaker independent KAIST continuous speech database (Park *et al.* 1995) is used for both training (89 speakers, 8790 utterance tokens) and test (21 speakers, 2073 utterance tokens). The test data set is composed of completely unseen utterance tokens in the sense that there is no overlap of either speakers or individual utterance tokens between test and training sets (see section 3.2 on page 50 for more details on data). The period between each parameter vector is assigned to be 10 msec and the size of analysis window is 25 msec.

As the main purpose of the experiment is to investigate results from acoustic probability, less effort has been made to refine the grammar model for phone sequence collocational restrictions. After all, language models in phone recognition do not play as a crucial part as in continuous word recognition. Results in Table 2.6 show that simple probability based language models enhance performance only slightly. Especially in Korean, where no syllable initial or final consonant clusters are allowed, phonotactic constraints are not very useful for narrowing the search path of consonants. Nor is the insertion of phones suppressed by imposing penalty scores on inserted items, as doing so can mask the basic effect of acoustic evidence while there may be a slight increase in accuracy.

Recognition accuracy for each broad natural class of phones is summarised in Table 2.6. In general, consonants seem to have higher recognition accuracy than vowels. But this does not imply that consonants are easier to recognise. As will be described in section 3.3 on page 51, I regard each diphthong as a single separate phone instead of a two phone sequence composed of a *glide* and a *vowel*. Consequently, the number of vowels has increased by 8, making a total number of 18 rather than the conventional number of vowel phonemes of Korean, 8 (Huh 1991, 183-184) or at most 10 (Lee 1996, 52). This narrow separation of vowels is thought to have caused the relatively poor vowel recognition.

It is consistently shown that obstruents other than fricatives are poorly recognised in comparison to sonorants. Especially, the recognition of stop sounds is not as good as other broad class categories shown, while accuracy of affricates is a little better than that. It should be noted that the accuracy of fricatives is the best of all the categories. This suggests that performance improvement for stops and affricates, among obstruents, is more badly needed than for other consonants, validating the current research which is targeted at enhancing the recognition of stop and affricate sounds. In a phone recognition experiment, Eun *et al.* (1989:110) also shows that the recognition of Korean stop sounds (accuracy 69.77%) needs to be further improved compared with sonorants (accuracy 75%) or vowels (accuracy 94.1%).

2.7.1 Error analysis

The obstruent confusion matrix of phone recognition in Table 2.7 is given to demonstrate the type of errors. It is shown that a large portion of errors are caused by failure in recognition of correct manner of articulation. In case of stops, place confusion is another major source of errors in addition to manner confusion. For example, the bilabial lax stop [p] is quite often recognised as an alveolar lax stop [t], or a velar lax stop [k]. But the amount of such place confusion is usually not as much as the manner confusion as compared in Table 2.8.

	c	c'	c^h	k	k'	h^h	p	p'	p^h	t	t'	t^h	s	s'	h	Sonorants	Vowels
c	2159	110	328	30	1	2	10	0	5	110	1	21	187	40	4	31	71
c'	13	193	11	1	0	1	2	0	0	0	4	2	0	0	0	1	4
c^h	36	26	418	0	0	4	0	0	3	1	0	12	9	0	1	2	10
k	79	7	24	2354	194	359	116	2	32	202	10	19	13	0	48	282	138
k'	3	2	0	41	863	102	3	1	4	6	23	9	0	0	1	0	6
k^h	3	1	9	51	22	354	5	1	6	4	2	8	4	0	22	4	6
p	8	0	1	76	1	4	778	24	175	77	12	12	5	2	22	79	83
p'	0	0	0	0	0	0	1	61	6	1	4	1	0	0	0	0	1
p^h	1	0	0	9	0	5	41	37	589	7	3	19	7	1	20	8	9
t	36	5	4	137	3	12	64	2	64	2219	50	235	26	1	46	86	67
t'	0	1	2	1	1	4	2	11	3	4	262	17	1	0	0	1	1
t^h	1	3	3	3	0	5	10	1	25	13	32	282	1	1	5	11	6

Table 2.7: Phone confusion matrix. Each item in rows is a reference phone and each item in columns is a recognised symbol.

Phone	Manner confusion (%)	Place confusion (%)
p	33.61	25.84
p'	71.43	26.67
p^h	46.15	14.20
t	34.01	23.99
t'	42.86	24.49
t^h	36.29	24.19
k	35.47	20.40
k'	71.14	11.94
k^h	48.67	9.33
Total	37.98	21.10

Table 2.8: Proportion of substitution errors caused by manner confusion compared with those of place confusion. Values are the ratio of each type error to total substitution error. Manner confusion refers to the number of phones recognised as either of the other two different types of the same place category (eg., [p] as [p'] or [^{h}p]). The place confusion refers to the number of phones recognised as difference place with the same manner (eg., [p] as [t] or [k]).

One noticeable type of error is that of lax stops. They are often recognised as various other sounds like sonorants and even vowels, while tense and aspirated sounds are much less likely to be realised as such. This fact reflects that lax sounds have relatively more allophones than the other two categories of sounds (see Table 2.2 on page 11). Especially intervocalically, where stops usually occur as voiced variants, their acoustic distance from the original voiced sounds like sonorants and vowels diminishes, causing relatively frequent confusion.

Affricates are less likely to be recognised as stops of different place because of the articulatory difference. Instead, they are found to be more frequently confused with alveolar fricatives [s, s']. However, this confusion is also distinctly less than three way manner confusion, as shown in Table 2.9.

Phone	Manner confusion (%)	confusion with sibilants (%)
c	45.96	23.82
c'	60.00	0
c^h	57.41	8.33
Total	47.59	21.44

Table 2.9: Proportion of substitution errors caused by manner confusion compared with confusion with sibilants [s, s']. Values are the ratio of each type error to total substitution error. Manner confusion refers to the number of phones recognised as either of the other two different types of the same place category (ie., [c] as [c'] or [^{h}c]). The other confusion refers to the count of each affricate recognised as sibilants.

In conclusion, errors caused by difficulty in manner distinction of stops and affricates turn out to be one of the main causes of phone recognition degradation, which need to be better taken care of.

2.8 Frequency of Stops and Affricates

It could be questioned how frequently the 12 Korean stops and affricates appear in utterances. Even if the segmental F0 model is found to be a useful way of distinguishing members of a stop class in laboratory data, introducing it into a recognition system might not give a noticeable improvement if the relevant phones are distributed too sparsely in a real data corpus. In order to see if the occurrence of the relevant sounds is frequent enough for influencing word and utterance recognition, I present the analysis results for some text data sets of Korean. The analyses are mainly focused on a specific position, that is, the initial position of a syllable or a word where the segmental F0 is believed to be most effective.

2.8.1 Data

In this kind of text analysis, one should be cautious in choosing data. The data sets collected for traditional phonetic experiments are unsuitable irrespective of size, since they are usually designed on purpose either to contain only part of the phone sets, or to include all the individual phones evenly distributed. It is equally inappropriate to analyse the texts of task oriented data which are artificially constructed for modelling speech characteristics. These do not reflect the distribution of phones in real spoken language, either, because the constructors of the database are likely to have the intention to cover the analysis units as evenly as possible. Therefore, I use three data sets none of which has been developed with such intentions stated above.

The first data set (Dataset-I) is a list of 2350 Korean written syllables which is registered as the ISO (International Standardisation Organisation) graphic character code set for Korean information interchange, under the name of KSX-0001(or formerly KSC-5601)[6]. Most modern Korean written and spoken words can be expressed in terms of those 2350 syllables[7].

The second data set (Dataset-II) is a word dictionary. The dictionary is constructed and provided by Natural Language Processing Laboratory, Pohang University of Science and Technology (POSTECH, http://nlp.postech.ac.kr). It comprises roughly 100,000 Korean words and morphemes. As this dictionary is not originally developed for speech systems but for natural language processors, only the written text of each word, without any information on pronunciation, is given.

Though the above two data sets are useful to understand the overall distribution of the stop and affricate segments, they do not constitute direct evidence about the spoken language. Thus, it is necessary to investigate the texts of real speech data. The third data set

[6]The document is obtainable from http://www.itscj.ipsj.or.jp/ISO-IR/149.pdf.

[7]Korean syllables such as 'pheung' or 'ttom' are not included among them but their appearance in real text or speech is extremely rare. For instance, a 100,000 word-level Korean dictionary described in the following paragraph does not contain any word containing those rare syllables.

(Dataset-III) is used for this purpose. It consists of 86 telephone dialogues and is about three hours long. Three hotel operators and 86 guests are involved. In each dialogue, a guest inquires about the availability and prices of hotel rooms, confirms a ready-made reservation, or makes a reservation. The whole data set is composed of 5561 utterances. The total word count is 16568, but 5802 different words are used including vocatives, interjections, and every kind of disfluency. As neither hotel operators nor guests are given any previous information about the experiment and the recording was done through tapping of the telephone line[8], all the utterance tokens are spoken in a completely spontaneous manner. Consequently, the style of speech is highly colloquial as well as purely natural. All the dialogues are transcribed by native Korean university students and errors are corrected through two-time cross verification of all the transcribed texts.

2.8.2 Methods and results

All the transcriptions, written in Korean, are converted to Roman characters for easy processing afterwards. Then, phone strings of each syllable and word are generated by automatic letter-to-sound mapping, taking into account most phonological and morphophonemic processes detectable through the transcribed text. For Dataset-III, as all the utterances are spoken spontaneously, a number of non-important expressions of disfluency are contained together such as hesitation, stuttering, exclamation, or trivial interjections. As they usually do not play a crucial part in the recognition of word or sentence meaning, they are excluded in counting.

Table 2.10 shows the distribution of each stop and affricate segment at the initial position of syllables and words. In all three data sets, more than half of the items begin with one of the stops or affricates.

Syllables and words with lax sounds occur more often than the other two classes, but aspirated sounds also appear with considerable frequency. The proportion of tense sounds is relatively small and this suggests that the distinction between lax sounds and aspirated

[8]The data set was used after obtaining the permission of each speaker.

Phone	(a) Syllables	(b) Dictionary Words	(c) Dialogue Words
p	5.5	8.21	6.25
t	5.4	5.22	5.22
k	7.3	12.73	15.78
c	5.7	10.65	13.06
LAX Total	23.90	36.81	40.31
p'	3.1	1.49	0.16
t'	3.7	1.19	1.69
k'	5.1	2.54	0.37
c'	3.5	1.28	0.20
TENSE Total	15.40	6.50	2.42
p^h	4.4	3.10	3.19
t^h	4.5	2.29	1.51
k^h	4.5	1.33	0.99
c^h	4.8	4.37	3.09
ASPIRATED Total	18.20	11.09	8.78
Total	57.20	54.40	51.51

Table 2.10: Proportion of syllables and words beginning with each stop or affricate. Column (a) results are calculated from Dataset-I (2350 syllables), Column (b) from Dataset-II (113194 dictionary words), and Column (c) from Dataset-III (82 natural continuous speech dialogues with 13063 linguistically meaningful words).

sounds is practically important. It should also be noted that the proportion of words beginning with affricate sounds [c, c', c^h] is fairly large, suggesting the importance of recognising affricates as well as stops.

In Dataset-II (column (b)) and III (column (c)), if the appearances of obstruents including those at word medial position (but still syllable onset position) are also counted, the proportion of relevant words increases to as much as 88.71 % and 85.96 %, respectively). This means the vast majority of Korean words contain at least one stop or affricate sound at a syllable onset position.

To summarise, a large portion of Korean syllables and words have at least one stop or affricate sound at their initial position so that correct recognition of those sounds, in either isolated word or continuous speech data, is expected to improve the performance of a recogniser.

2.9 Limitations of Temporal Cues

As indicated by Liberman (1996), both many-to-one and one-to-many relations exist between acoustic cues and phonetic contrasts. Multiple phonetic contrasts can be expressed by a single acoustic cue while a single phonetic contrast can be defined by multiple acoustic cues. Focusing on the current subject with the latter aspect, the manner of Korean obstruents can be classified in terms of more than one single acoustic cue. In other words, segmental F0 perturbation is by no means the only cue for distinguishing the three manners of articulation. Earlier phonetic studies have suggested various correlates for manner distinction of Korean stops and/or affricates such as *intensity build-up* at the vowel onset (Han & Weitzman 1970), *glottal width* in consonant articulation (Kagaya 1971), *airflow rate* (Hardcastle 1973), as well as temporal characteristics (literature given below). As it is the temporal characteristics which have been most frequently suggested as the principal cue in much literature, further examination is necessary with a view to indicating their limitations and supporting the necessity of using F0 cues.

Though some studies report the durational properties of vowels depending upon the type of neighbouring consonants (Mohr 1971, Zhi 1993), more studies have concerned the durations of consonantal parts. It has been observed that there are at least two separate cues contributing to stop classification: voice onset time (VOT) and duration of occlusion. A number of earlier studies on Korean phonetics have examined the VOT difference (Lisker & Abramson 1964, Abramson & Lisker 1971, Hardcastle 1973) and more recent experiments have included closure duration as another cue (Zhi *et al.* 1990, Silva 1992, Han 1996, J. Pae & Ko 1999, Yun & Jang 1999). General agreement among them is that tense stops have the longest closure period and lax stops have the shortest closure

period whereas aspirated stops have the longest VOT and tense stops have the shortest VOT regardless of place of articulation, which can be summarised as:

Closure Duration TENSE > ASPIRATED > LAX

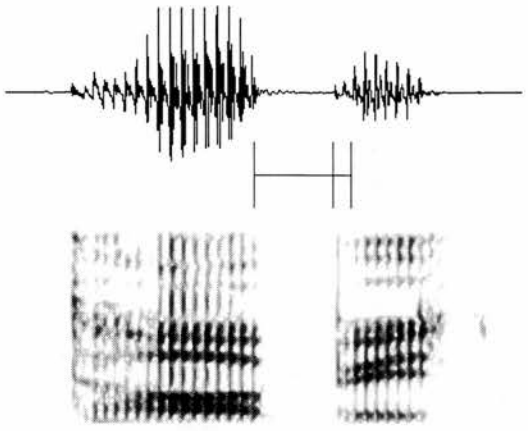
Voice Onset Time ASPIRATED > LAX > TENSE

Figure 2.4 is given for the visual comparison of those durational differences shown in the speech of bilabial stops of each class.

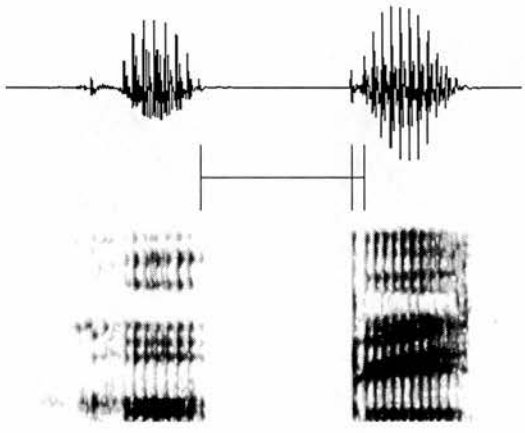
In spite of those findings on durational cues, I would like to point out their limitations of use, in an effort to validate the necessity of another cue, vowel F0.

First of all, use of the temporal cues is highly dependent upon the context of the consonant. When a stop sound is preceded by any kind of silence, closure duration is unmeasurable since the articulatory occlusion is not separated from other types of silence (see the beginning part of each picture in Figure 2.4). Consequently, the durational effects are considered to be better observable when the consonant is located intervocalically, or at least surrounded by voiced sounds. However, in those positions, the necessity of the distinction, either by duration or F0, is rather reduced since the lax stops quite frequently undergo a change into corresponding voiced consonants, homorganic fricatives[9], or sometimes to nearest approximants. Figure 2.5 is a typical example of a voiced bilabial intervocalic stop to be compared with non-voiced version Figure 2.4 (a). In the case of voice assimilated lax stops like this, acoustic distance between lax stops and other stops, that never realise as a voiced counterpart, becomes greater due to the voicing difference, so less confusion is expected between them. This means that temporal cues may not be necessary for lax stop identification even at the intervocalic position where they are supposed to work better.

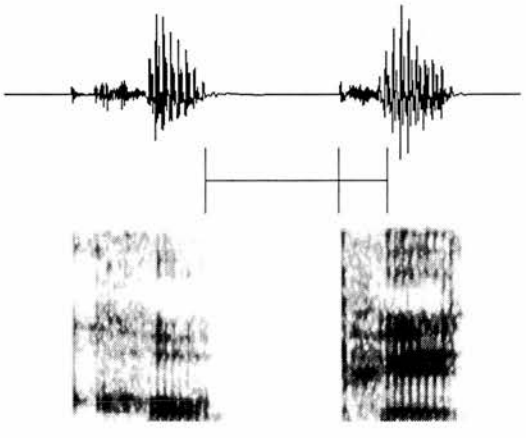
[9]Namely, bilabial fricatives [ɸ, β] for bilabial lax stop /p/ and velar fricatives [x, ɣ] for velar lax stop /k/. But the alveolar lax stop [t] is not likely to undergo the change into its homorganic fricative [s, z]. One possible explanation is that those segments already occur as a phoneme (/s/), or its allophone ([z]), and weakening of /t/ into one of those fricatives is discouraged with a view to keeping recoverability by avoiding merges at the phonetic level.



(a) CV p i



(b) CV p' i



(c) CV ph i

Figure 2.4: Comparison of closure duration and VOT of intervocalic bilabial stops. Words are (a) *teo-pi* 'race', (b) *ko-ppi* 'reins', and (c) *kho-phi* 'nosebleed', all spoken by a single male speaker in the same recording session. Two marked intervals in each token represent closure duration and VOT, respectively.

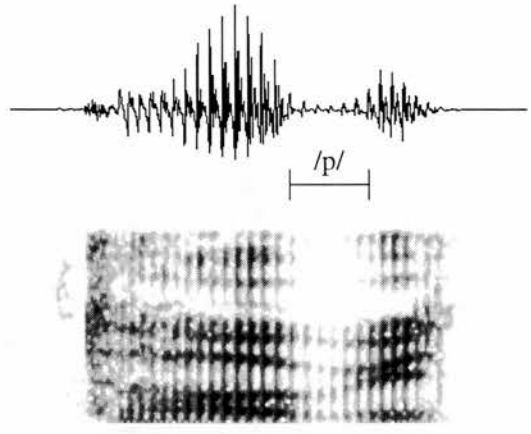


Figure 2.5: An example of weakened intervocalic stop. Word spoken is *teo-pi* 'race', different utterance of the same word as in Figure 2.4 (a), spoken by the same speaker. The manifest appearance of formants and voice bar in the spectrogram suggests that the phoneme /p/ is phonetically realised as a voiced sound, like a bilabial fricative [β] or an approximant [w].

In the carefully designed phonetic studies from which most durational statistics are obtained, this problem has been avoided on purpose by constructing data in such a way that those weakened intervocalic stop tokens are not produced. In the natural connected speech which most current research on ASR is targeting, however, such weakening happens quite frequently so that it is less plausible for temporal cues to play an important part.

In brief, duration is not as dominant a recognition cue as it has been considered and claimed to be, when it comes to the analysis of connected speech data rather than controlled laboratory speech.

Another drawback of the temporal cues is that they are mainly effective for stop sounds, but not as good for affricates. It is obvious from the phoneme inventory of Korean (see section 2.2) that the three-way distinction of manner of articulation is available for the affricate sounds as well as stop sounds. In fact, affricates are similar to stops in various

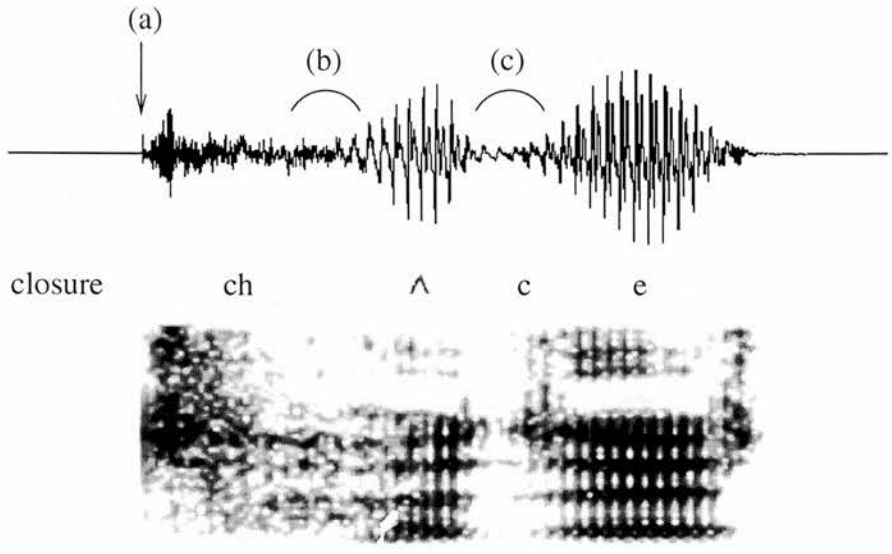


Figure 2.6: Waveform and spectrogram of *cheo-ce* [cʰʌce] 'sister-in-law'.

ways, together composing the natural class which is represented with phonological distinctive feature [-continuant], also existing only as voiceless in the underlying level. But the crucial difference occurs in the later stage of the articulation. While the occlusion of the airflow takes place in the similar way to stop articulation, plosion of affricates is not as abrupt as that of stops. Figure 2.6 is an illustration of how word initial /cʰ/ and intervocalic /c/ are phonetically realised. In word initial position, the boundary between occlusion and frication noise, marked as (a) in Figure 2.6, is shown as distinctly as in stop sounds. But as mentioned already, the closure duration is not observable at this position because of the preceding silence. As for VOT, it is quite difficult to demarcate the position where the affricate ends and the following vowel begins, due to the superimposition of frication onto vowel onset period (Figure 2.6 (b)). As a consequence, the measurement of VOT is hard as well.

It is also shown that the weakening process which frequently changes the identity of stops applies to the lax affricate sound (/c/) as well at the intervocalic position (Figure 2.6 (c)). The segment is phonetically realised as voiced fricative sounds [z] or [ž], without any

distinct trace of occlusion. This means that temporal cues are not much of use in this position for affricates as well as stops.

This restriction does not exist in the F0 effect at least at the word initial position. The statistical results shown in the next chapter confirm that the distribution of vowel F0 after affricates is as consistent and uniform as after stop consonants. This makes it possible to treat affricates in line with stop sounds.

My intention in this section is not to deny the usefulness of the temporal characteristics altogether. As earlier studies have shown through perception and production tests, duration is certainly the most commonly considered acoustic cue for manner identification of Korean obstruents. Han (1996:178) also mentioned that "... even though F0 is a highly significant perceptual cue in distinguishing between tense and plain consonant, it does not appear to be the only cue ...". For example, VOT is still a useful cue for all three stop classes at the word initial position. Moreover, both VOT and closure duration are useful for at least two classes such as tense and aspirated stops at the intervocalic position, where segmental F0 is less useful.

What I would like to point out is the flexibility of multiple cues in accounting for a single linguistic contrast. Even if it is generally the case that temporal characteristics are overall the principal cue for obstruent manner differentiation, the priority can be changed depending on the different segmental or prosodic environments. A noteworthy comment related to this is made by Skaličková (1959).

> The main or what we call the primary property of a sound need not maintain its priority under all circumstances, but under certain conditions it may be replaced by another, originally secondary property.

Where the temporal cues are not quite effective because of various reasons given above, vowel F0 becomes more active, playing a compensatory role in identifying the manner of articulation of obstruents. In this sense, it is not coincidence that the segmental F0

effect is most outstanding in such contexts as prosodic phrase boundary initial positions and post-pausal positions, where temporal effects are quite feeble.

Before ending this section, the difficulty of implementing duration models is worth mentioning. The traditional Hidden Markov Model is not capable of introducing duration of units properly, as the probability of a unit remaining in the same state is exponentially proportional to the elapse of time as follows:

$$P(\tau|q_i) = (a_{ii})^{\tau} - 1(1 - a_{ii})$$

where $P(\tau|q_i)$ is the probability of remaining in state q_i for τ time units;

and a_{ii} is the self-transitional probability.

To overcome this difficulty more explicit implementation of the duration model has been proposed (Levinson 1986, Ferguson 1980), but it remains to be seen that these methods will also be useful for the recognition of Korean stops and affricates.

The reliable extraction of the durational characteristics is also difficult. The difference between the durations of stop sounds for each class is highly variable depending upon speech rate of the utterance. The faster the speech, the less distinguishable. Therefore, it is harder to extract a consistent characteristics from the spontaneously spoken speech which most modern ASR systems work on. Besides, the frame rate at which a signal is analysed can undermine the statistical significance of durational difference. For example, if the frame length of digital signals is set up as 10 msec, as is the case in many speech analyses, the minimal VOT value extractable from a signal is also 10 msec at the shortest. But it is often the case, especially in natural connected speech, that Korean tense stops have shorter than 10 msec and lax stops, a little longer than 10 msec. Then it will be hard to detect the VOT of tense stops and harder for an automatic recogniser to distinguish tense stops from corresponding lax stops in terms of the VOT cue due to the inevitable rounding-up error.

2.10 Summary

Segmental and prosodic structures of Korean have been introduced as a background for the current research. We have also seen how the human production mechanism generates three different consonant classes and how the pitch is affected at the onset of vowels following each class.

Empirical and theoretical support has been given to the study of Korean non-continuant obstruents in automatic speech recognition. First, the restricted segmental structure within syllables permits vowels to maintain post consonantal effects relatively easily. Second, thanks to the fact that the other factors affecting vowel F0 have their influence mainly at the right edge of the prosodic phrases, segmental F0 effects survive utterance level F0 fluctuations, without being completely obliterated. Third, recognition accuracy of stops and affricates is not as good as other consonants or vowels in continuous speech data. Fourth, a large proportion of Korean syllables and words contain stop or affricate sounds especially in their initial position where segmental F0 perturbation is supposed to be most effective. And finally, the temporal characteristics which have been widely accepted as a principal cue for stop consonants have limitations which the vowel F0 cue does not have. There seems to be a dynamic interaction between duration and F0 in such a way that one is able to carry phonetic distinction when the other is not.

In conclusion, it is worth further investigating the behaviour of vowel F0 perturbation and exploiting it for speech recognition.

CHAPTER 3

Speech Recognition

3.1 Introduction

In this chapter, I will describe how I established the baseline speech recognition system based on phone units.

This baseline model is utilised in many ways throughout the experiments in this thesis. Two primary ones are as follows. First of all, it plays a role of benchmarking. Any further enhancement of recognition performance, such as pronunciation modelling, demisyllable unit recognition, as well as segmental F0 implementation, will be compared with this baseline performance. Second, it is used as an automatic phone labeller. Phone labels are necessary in the statistical analysis to verify the segmental F0 perturbations and useful in modelling pronunciation variation described in the later part of this chapter.

Though the major aim of this chapter is the baseline model construction, various techniques used in doing so will continuously be employed throughout the recognition experiments in the thesis after necessary modifications.

The chapter is composed of two major parts. First, the development of a context dependent phone-based recogniser through a Hidden Markov Model training technique will be

described. Second, an attempt is made to improve recognition performance by lexicon expansion. A technique of modelling pronunciation variation adopting both knowledge-based and data-driven approaches will be introduced.

3.2 Data

The major data set used for various connected speech recognition experiments in this thesis is one of five data sets constructed and provided by the *Communications Research Laboratory* of the *Korea Advanced Institute of Science and Technology* (KAIST), which I will call the 'KAIST data(base)' (Park *et al.* 1995).

The data set is composed of utterances used in trade negotiation so that the vocabulary contains many trade related terms including company names. Many words are said to have been selected from a conversation textbook of the same task domain but new words were also added in order to generate natural utterance tokens. The total vocabulary size is 2920, a number which could vary a little depending on how the term *word* is defined (see page 54). Park *et al.* (1995) states that sentences in the recording scripts were artificially designed to even out word frequencies, and that combination of words into sentences was designed to produce as natural sentences as possible. The average number of words in a sentence is 8.4.

In the original database, a total of 150 speakers participated in the recordings, but I only used 110 speakers' speech, excluding utterances of speakers from a few specific regions because the prosodic structure of their dialects is known to be quite different from the target dialect, Seoul Korean, generally accepted as the standard version of Korean. For example, Kyeong-sang dialect has lexically contrastive tones, which will seriously mask the segmental F0 effect. Most speakers are in their twenties or thirties with university education. Each speaker uttered 90-100 sentences and there are a total of 10863 utterance tokens. I divided the data into two subsets: one for training, the other for test. Below are details on each subset.

	Training Data	Test Data
Speaker	89 (35(f) + 54(m))	21 (9(f) + 12(m))
Utterance Tokens	8790	2073
Vocabulary Size	2920	
Dialect Regions	Seoul, Kyeong-ki, Kang-weon, Chung-cheong, Je-ju	

As no speakers or tokens in the training data also appear in the test data, the two data subsets are completely exclusive of each other. Recording was conducted in a silent, but not sound processed, office and raw data signals are digitised via 16-bit quantisation at a 16000 Hz sampling rate. Orthographic transcriptions for sentences are also provided, but no information on word or phone boundaries is given.

3.3 Phones as Basic Units

The basic units of the baseline system are phones which are the most frequently used units of modern ASR systems thanks to the compact number of units and their applicability for context dependency.

The number of Korean phonemes varies a little from one scholar to another, mainly due to disagreement on the number of vowels. But it is generally agreed that there are 19 consonants, 8-10 vowels, and 3 approximants. I use all the consonants as recognition units. However, some modification is done for vowels and approximants. Instead of treating approximants as individual units, I regard them a part of the vowel. Accordingly, each of the diphthongs is regarded as a single unit as with other Korean ASR systems using phone like units (eg., Yun *et al.* (1997), Kiecza *et al.* (1999)). In fact, either way may not be crucial to system performance eventually, since recognition techniques such as parameter tying and clustering can produce similar effects. However, it is for the convenience of practical processing of data that I treat diphthongs like single vowels. For example, visual determination of approximant boundaries is extremely difficult in hand labelling because of the dynamic nature of their articulation. In Korean, as the structure

of diphthongs is usually 'approximant + vowel' (eg., /ya/, /we/), the approximant is quite often deleted or the duration is generally short. Thus, separation of approximants from vowels may cause difficulty in training and the recognition those sounds. After all, speech recognition is essentially phonetic, and its units do not have to be phonologically defined.

The phone units I use for building recogniser are as follows (see Appendix A for corresponding IPA transcription):

(3.1) **Korean Phone Units** (total 34)

- 15 obstruents: p t k c p' t' k' c' p^h t^h k^h c^h s s' h
- 4 sonorants: m n ng l
- 8 monophthongs: a e i o u v x
- 10 diphthongs: wa we wi wv xi ya ye yo yu yv
- 1 silence: sil

3.4 Hand Labelling

Phonetic boundaries marked by hand are useful in various ways. Above all, it helps make a good initial estimation of acoustic model parameters in the HMM training phase. Specific boundary information is not strictly necessary for statistical approaches to speech recognition as long as word level transcriptions and word lexicon with phone strings are given. Nevertheless, providing a certain amount of accurate time information makes it possible to save time and data necessary for constructing a reliable system. This sort of bootstrap training will be used through the recognition experiments in this thesis.

Second, phone level annotation is also used for labelling higher level constituents such as syllables and demisyllables which are crucial for implementation of the segmental F0 as described in the later chapters. These higher level labels are extractable automatically once phone labels are given and proper syllabification criteria are established.

Third, another use of phone labels is in the phonetic verification of the segmental F0 effect. As already mentioned it needs to be confirmed that this effect is still observable in data which have not been collected for the specific purpose of observing the segmental F0 effect. Phone labels are indispensable for this purpose along with appropriate F0 extraction.

Fourth, the difference between phonemic strings and the phonetic labellings will reveal the phonological rules of a language. When the rules are established, pronunciation variants can be automatically generated and used effectively for pronunciation modelling and subsequently for the lexicon. This procedure will be described in detail in section 3.9 of this chapter.

Phone-level phonetic labelling was performed by hand over 433 sentence tokens. Those sentences are selected to include all the phones as evenly as possible. For example, tokens containing relatively rare phone units like [p', t', k', wi, ye] are chosen in the first place. The estimate of the total duration of the labelled data is approximately 21 minutes and average duration of each token is 2.875 sec. I and another Korean (Weonhee Yun), who is also working for a postgraduate degree in the speech related area, did the labelling making use of the tools *xwaves* and *xlabel* of *Entropic* (Entropic 1998).

Basically phonetic labelling rather than phonological labelling is adopted. By phonetic labelling, I mean that phones instead of phonemes are used as basic label units. This is to provide as detailed and practical information as possible. But allophonic variations which are predictable from context are not marked. For example, alveolar liquids /l/ and /r/ are labelled identically since they can be easily relabelled automatically considering that /r/ segments appear only intervocalically[1]. Regarding stop consonant marking, closure duration and aspiration period are separately specified so that each of them can be used

[1] Strictly speaking, this is not the case as there seem to be a few counter examples. I found several cases where /l/ pronunciation is still maintained intervocalically. But I assume they are not standard pronunciation of Korean.

to find durational characteristics. When stop sounds are preceded by silence an arbitrary length (approximately 50msec) is given as stop closure.

As hand labelling is tedious and error-prone work, label files need to be verified before being seriously used. Though there is no complete method for detecting errors automatically, some of them can be found using the information provided by phonotactic constraints. For example, when a stop closure marking is found preceding a non-stop sound it must be a mislabel. After correcting such obvious errors, human visual inspection is conducted for the entire set of labels to ensure the absence of major errors.

3.5 Word Decomposition

The target of recognition in typical modern recognisers is the *word*. Consequently, relevant subsystems are constituted around this unit: a lexicon is composed of word items, a language model is to constrain cooccurrence of words, and recogniser evaluation is usually in terms of word accuracy or word error rate. For the convenience of processing in the recogniser construction, the formal abstract definition of the term *word*, such as 'a minimal communication unit', is often assumed to be identical with a practical definition like 'an item surrounded by white spaces in the text'. In languages like English, this assumption is available as the two types of definition are normally in agreement. But more elaboration is necessary for the languages, like Korean, in which multiple morpheme agglutination is a frequent method of word formation. A word stem is usually attached by one or more bound morphemes to form a single communication unit. Table 3.1 illustrates some examples. There are more than 10 such suffixes which can be attached to a single noun either on their own or even in combination.

The convention for constituting a lexicon in most Korean speech recognition systems is to have all the morphologically derived items with the same stem specified in the lexicon. For instance, each of the eight items in Table 3.1 is regarded as a separate lexical item. As a consequence, the vocabulary in a system can become large very quickly. In a small

word item	structure	glossary
na	STEM	'1st person singular'
na-neun	I + TOP	'I (subject) ...'
na-eke	I + DAT	'to me'
na-man	I + RES	'only I' or 'only to me'
na-eke-neun	I + DAT + TOP	'I (possess) ...'
na-eke-man	I + DAT + RES	'only to me'
na-man-eun	I + RES + TOP	'only I ... (negative)'
na-eke-man-eun	I + DAT + RES + TOP	'at least only to me'

Table 3.1: Example of Korean morpheme agglutination. The grammatical markers used are: TOP(ical), DAT(ive), RES(trictive).

vocabulary recognition system, this may be tolerated. For a system with a moderate or large size vocabulary, it begins to be a problem. Vocabulary independent systems can hardly be established in this way since the increase of lexical items will be immense if all the combination of 'stem+affix(es)' is specified in the lexicon. In my baseline system construction, I will separate such grammatical morphemes (eg., *neun, eke, man* in Table 3.1) from stem words and specify them separately in the lexicon. Subsequently, the language model is also developed to reflect this separation. The vocabulary size, 2920, of the current KAIST data is the count after this treatment. Otherwise, the size would increase to approximately 3200. The extent of reduction is not enormous in this case, but the effect will be more prominent in a large vocabulary lexicon.

It should be mentioned that morpheme separation is conducted only for words and affixes with respect to noun categories. The Korean verbs or adjectives are also highly agglutinative but separation is not attempted for them mainly because of the extremely complicated morphological or phonological processes which intervene when each grammatical affix is attached to a stem of those categories. More sophisticated morphological and phonological analyses are necessary to resolve this problem, which is beyond the scope of the current research. Owing to this restriction, it is not appropriate to call the current approach a genuinely morpheme based approach.

3.6 Basic Lexicon

To the best of my knowledge there has been, before now, no publicly accessible pronunciation dictionary for Korean speech recognition. Fortunately, however, Korean pronunciation is roughly predictable on the basis of orthography though some words require more concrete morphological or even etymological information for correct generation of pronunciation. A baseform lexicon with one canonical pronunciation for each word item is automatically generated by the following steps:

(3.2)

- Romanisation
- Letter to phoneme conversion
- Phonological rule application

Romanisation

As the first step, Korean characters are mapped into corresponding English alphabet characters. This means that each 2-byte 8-bit syllable character is converted into one or more 1-byte 7-bit phone characters. This conversion is not indispensable since the direct mapping from Korean characters into a phone string is not impossible. But using machine readable anglicised forms throughout the recognition process is a lot more convenient for various purposes like software compatibility and text readability. In fact, most Korean word and sentence examples demonstrated in this thesis are also the output of this conversion. The detailed description of Romanisation is given in Appendix B on page 158.

Letter to phoneme conversion

Each character is mapped into a corresponding phoneme according to a simple one-to-one mapping criterion so as to produce the underlying phonological representation.

	Phonological Rule	Conversion Example	
a.	Cluster simplification	k a p s → k a p	'price'
b.	Coda neutralisation	i p^h → i p	'leaf'
c.	Consonant nasalisation	s i p yu k → s i m n yu k	'sixteen'
		k u k m u l → k u ng m u l	'soup'
d.	Tensing	s u k c e → s u k c' e	'homework'
e.	Aspiration	s i l h t a → s i l t^h a	'hate+END'
f.	Palatalisation	k u t i → k u c i	'obstinately'
g.	Laterallisation	k^h a l n a l → k^h a l l a l	'blade'

Table 3.2: Phonological rules for generating canonical pronunciation from underlying phonological representation.

Phonological rule application

As the final step for establishing the lexicon, various phonological rules are applied to the underlying structure to generate one canonical pronunciation for each word item, which is illustrated in Table 3.2.

The phonological rules employed here are the ones widely agreed by phonologists of Korean[2]. The important common nature of the rules used here is that the rule application is obligatory or nearly-obligatory, irrespective of speaker identity, or speech rate. In other words, failure to apply these rules will generate unacceptable pronunciations.

Generating phone strings from orthography through letter-to-phone mapping and phonological rules is not always possible. For example, multi syllable words derived from Chinese characters are sometimes not subject to the above phonological rules. Some Korean words also have a unique canonical pronunciation which cannot be predicted by rule application. These ad-hoc words are included in the lexicon after separate manual treatment.

[2]Some Korean scholars, following the tradition of structural linguistics, divide phonological rules into morphophonemic rules, which are mainly caused by morpheme adjacency, and allophonic rules, usually applied for the sake of pronunciation convenience (Huh 1991, Lee 1996). As the application of most morphophonemic rules is obligatory, the rules in Table 3.2 seem to be of this type. But I have no intention of following such a distinction as it is controversial in the framework of later generative phonology.

As a result, an initial version of the lexicon, with only one canonical pronunciation for each word, is constructed. A few lexical items are given below for demonstration.

kwi-sa	k wi s a
noh-chyeoss-seup-ni-ta	n o t ch yv s' x m n i t a
math-ko	m a t k' o
myeong-ham	m yv ng h a m
pu-sok-phum	p u s o k ph u m
ha-kess-seup-ni-ta	h a k e s' x m n i t a

I will call this the 'base lexicon', which will be further modified later (section 3.9) for improvement of recogniser performance.

3.7 Language Model

Constraining the sequence of recognised words is very important in speech recognition. Two different types of language models can be considered. High level grammar is the first option. This would directly use the syntactic structure of a human language based on a theoretical description. As it is a rule based approach, it has the advantage of being independent of task domain or vocabulary size. However, few modern speech recognisers have adopted this type of language model since its implementation in recognition systems is very difficult. In addition, describing the optimal grammar is another problem, taking into account the broken syntactic and semantic structures of spoken utterances. Thus, the alternative to a knowledge-based grammar is a statistical language model, which is adopted by most state-of-the-art ASR systems to constrain word sequences.

In the framework of statistical language modelling the problem of connected speech recognition can be formally represented as finding a word string satisfying

(3.3)

$$\hat{W} = \arg\max_{W} P(A|W)P(W)$$

where $P(A)$ denotes the probability of acoustic evidence and $P(W)$ is the language model probability.

Bigram and trigram are the most frequently used types of language models in modern ASR. In my experiments, I use only bigram language models because trigrams do not seem to improve the recognition performance with the current data. In a bigram language model, the estimate of the probability of a word sequence 'w_1 w_2' is represented by relative frequency, as

(3.4)

$$\hat{P}(w_1, w_2) = \frac{C(w_1, w_2)}{C(w_1)}$$

In language modelling, it is impractical to collect sufficient training data to cover all possible sequence of words. Accordingly, smoothing of data is necessary to take care of unseen or infrequent word strings which will appear in test data. Two well-established methods, *discounting* and *backing off*, are adopted in the current experiment.

Discounting

The probability estimation in terms of frequency counting in a training data tends to be biased toward observed data. That is, unseen events are assigned probability zero, even though some events that do not occur on the training data are very likely to appear in test data. To take this into consideration, redistribution of probability is useful such that a portion of observed event counts is allocated to counts of unseen events.

Thus the Formula 3.4 can be revised as

(3.5)

$$\hat{P}(w_1, w_2) = \frac{C(w_1, w_2) \times D}{C(w_1)}$$

where D is the discounting coefficient, whose value is between 0 and 1.

Among various ways to determine the discounting coefficient (see Clarkson & Rosenfeld 1997 for review), Ney *et al.*'s (1994) method called *absolute discounting* is adopted. The value D is calculated as

(3.6)

$$D = \frac{C(w_1, w_2) - b}{C(w_1, w_2)}$$

while the constant b is estimated by

$$b = \frac{n_1}{n_1 + 2n_2}$$

where n_r is the number of events which occur r times. That is, n_1 is the number of bigrams which appear only once, and n_2 is the number of bigrams which appear twice. Thus, b is calculated independent of w_1 and w_2.

This method is used simply because it empirically gives the best recognition accuracy. The *perplexity*, a useful measure of language model, of the bigram established by this method is also lowest for the current corpus (see below for more description of perplexity).

Backing off

In case a given bigram word sequence (w_1, w_2) has been observed less than a designated threshold (k) in the training data, $P(w_1, w_2)$ is replaced by the unigram probability $P(w_2)$

which is given by

(3.7)

$$P(w_2) = \begin{cases} \frac{C(w_2)}{N} & \text{if } C(w_2) > f, \\ \frac{1}{N} & \text{otherwise} \end{cases}$$

where f is a unigram floor count (I used 1 as its value), $N = \sum_{n=1}^{V} \max[C(w_n), f]$ (V is the vocabulary size and w_n is nth word in the vocabulary).

Finally, the bigram language model probability $P(w_1, w_2)$ for the recognition experiments through the thesis can be summarised as

(3.8)

$$P(w_1, w_2) = \begin{cases} \frac{C(w_1, w_2) \times D}{C(w_1)} & \text{if } C(w_1, w_2) > k, \\ \alpha P(w_2) & \text{otherwise} \end{cases}$$

where the back-off weight α is calculated to ensure that the sum of $P(w_1, w_2)$ for all the vocabulary words w_2 equals to 1. In the current experiment the back-off threshold k is fixed to 7.

Domain of language model calculation

It is reasonable to assume that the initial word of a syntactic domain at some level is not systematically related to the identity of the last word of the previous domain. I regard a clause as such a level in Korean. This boundary information is taken into account when language models are created. For example, if a sentence token is composed of two or more clauses, the domain of bigram formation is clause-internal, having the first and last word of a clause backed off to a unigram.

Evaluation of language model

Perplexity, a concept originated in *information theory* (Usher 1984), is widely accepted as an objective measure of language model quality. Perplexity is an intuitive expression of *Entropy* which is a measure of difficulty in each word based on the language model. Using an estimate of a word string probability $\hat{P}(w_1, w_2, ...w_n)$, the amount of information per word in a text corpus can be measured as (source: Jelinek (1990:474))

(3.9)

$$LP = -\frac{1}{n}[log\hat{P}(w_1, w_2, ..., w_n)]$$

while the perplexity(PP) is given by

(3.10)

$$PP = 2^{LP} = \hat{P}(w_1, w_2, ..., w_n)^{(-\frac{1}{n})}$$

For a proper evaluation of language model, the test text should be independent from the one used in building language model. Approximately 20% of the training data has been reserved for this purpose. The perplexity of the bigram calculated over the held-out data is 25.74, which can also be intuitively interpreted as the average number of possible words following a word. In the test text, there are 62 new words which do not exist in training text and the default probability is assigned to them in terms of the discounting and backing-off scheme.

This language model will be consistently used for every recognition test regardless of model type.

3.8 Training and Recognition

The Hidden Markov Model Toolkit (HTK, Young *et al.* 1996) is used for both training and test of recognisers.

Features

Features are extracted from waveform signals for each 10 msec frame by using 25-msec Hamming window with pre-emphasis coefficient 0.97. A 39 dimensional vector is allocated for each frame, which is composed of 12 Mel Frequency Cepstral Coefficients (MFCC), energy, and their first and second derivatives.

Type of models

A left-to-right 9 Gaussian-mixture continuous HMM with 3 emitting states is created for each phone unit. The output distribution for each frame in a state can be represented by a *multivariate Gaussian mixture density* function and its parameters are formally represented as follows.

(3.11)

$$\mathcal{N}(o : \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)' \Sigma^{-1} (o-\mu)}$$

where $\mathcal{N}(o : \mu, \Sigma)$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, and n is the dimensionality of observation o.

To reflect the fact that a brief pause can be located between two words, a short pause model with a single state, which is tied to the centre state of silence model, is separately created.

Training

In training, if boundaries of data tokens are explicitly marked, better initiation of parameters can be accomplished, which is called bootstrap training. The hand labels are used for this purpose. The HTK (program HInit) uses a *segmental k-means* algorithm (Rabiner & Juang 1993) followed by *viterbi* alignment for this initial parameter estimation. Then, the *Baum-Welch* algorithm is used for parameter re-estimation (HTK program, HRest) again using all hand labelled time information. After initiation, all training data are used for *embedded training* (HTK program, HERest) in which re-estimation of parameters is performed over all data in parallel by accumulating values of each data token, again by the *Baum-Welch* algorithm.

Context dependent models

The obvious drawback of the individual phone unit is the lack of capability to capture across-phone coarticulation. Context dependent phone modification is a widely testified solution, which I also applied to enhance performance. Phones are expanded to triphones and acoustically close triphones are tied together in order to alleviate data insufficiency. In doing so, I employed a method called *tree-based clustering* in which a decision tree, manually established on the basis of phonetic grouping, is used to classify the acoustically similar contexts of triphone boundaries.

Recognition and evaluation

A powerful pattern recognition technique known as *viterbi* algorithm (Forney 1973), which computes the most likely state sequence, is used for the recognition process (The program HVite in HTK).

To evaluate recognition performance, output from the recogniser is compared with the ready made word level transcriptions using string alignment procedure based on *dynamic programming* technique (Bellman 1957).

The word accuracy is calculated by taking into account the correctly recognised words against errors such as insertion, deletion, as well as substitution, while *word error rate* (WER) is simply the complement of accuracy, which is to say:

(3.12)

$$Accuracy(\%) = \frac{N - (Substitution + Insertion + Deletion)}{N} \times 100$$

$$WER(\%) = 100 - Accuracy$$

With the data and methods described in this chapter, the best word accuracy obtained is 78.47% with the bigram language model with perplexity 25.74. This result (WER 22.53%) seems to be reasonable compared with the performance of an English speech recogniser (King 1998:42) which was constructed in a similar way to the current system. Its WER was 24.8% and the bigram language model perplexity was 23.6. Though the database used in his system is a spontaneous speech corpus, compared with a read speech database currently used, its vocabulary size (900) is smaller than the current system (2920). Thus, an indirect comparison seems reasonable.

3.9 System Improvement Using Pronunciation Variation Modelling

The pronunciation of a given word varies. It is virtually impossible even for a single speaker to pronounce the same utterance twice in an acoustically identical fashion. The degree of variation increases with the the number of speakers. The KAIST data used for the current research is also expected to contain a good deal of pronunciation variability. Although the data tokens were created by reading ready prepared scripts, they were produced in a fairly natural fashion as the vocabulary and style appear to have been chosen

on the basis of conversational speech. Besides, there is a fairly large number of speakers (110), and speech style and speech rate of each speaker are also found to be quite variable.

From the recognition point of view, this variability appears to be one of the causes of errors. Research has shown that appropriate modification of the lexicon to model pronunciation variation can enhance the recognition performance significantly (Schiel *et al.* 1998, Ferreiros *et al.* 1998).

3.9.1 Overview

There are many different ways of modelling pronunciation variation, but they can roughly be classified into a *top-down* approach and a *bottom-up* approach. The top-down approach adopts a more knowledge-based method based on abstract linguistic studies. It uses the phonological representation and previously formalised rules to derive variants of pronunciation. The bottom-up approach, on the contrary, trusts more in what is concretely revealed by speech data. Pronunciation variants can be collected by direct observation of such data. I take advantage of both approaches. To obtain a base form for each word, I use phonological rules established by traditional linguistic studies. To generate as many variants as possible, I rely on other rules which I discovered by inspecting manual phonetic labels. Then, to select out only the practically useful variants, I investigate the automatic labels which are the products of the recogniser itself. The procedure is summarised as a block diagram in Figure 3.1.

3.9.2 Base lexicon with canonical pronunciations

The starting point for variation modelling is the base lexicon, which has been described earlier in this chapter. The phone strings in it are assumed to be the standard pronunciation for each word as they are derived in terms of major phonological rules, although this may be the case only when each word is pronounced in a citation form and in a relatively slow speech.

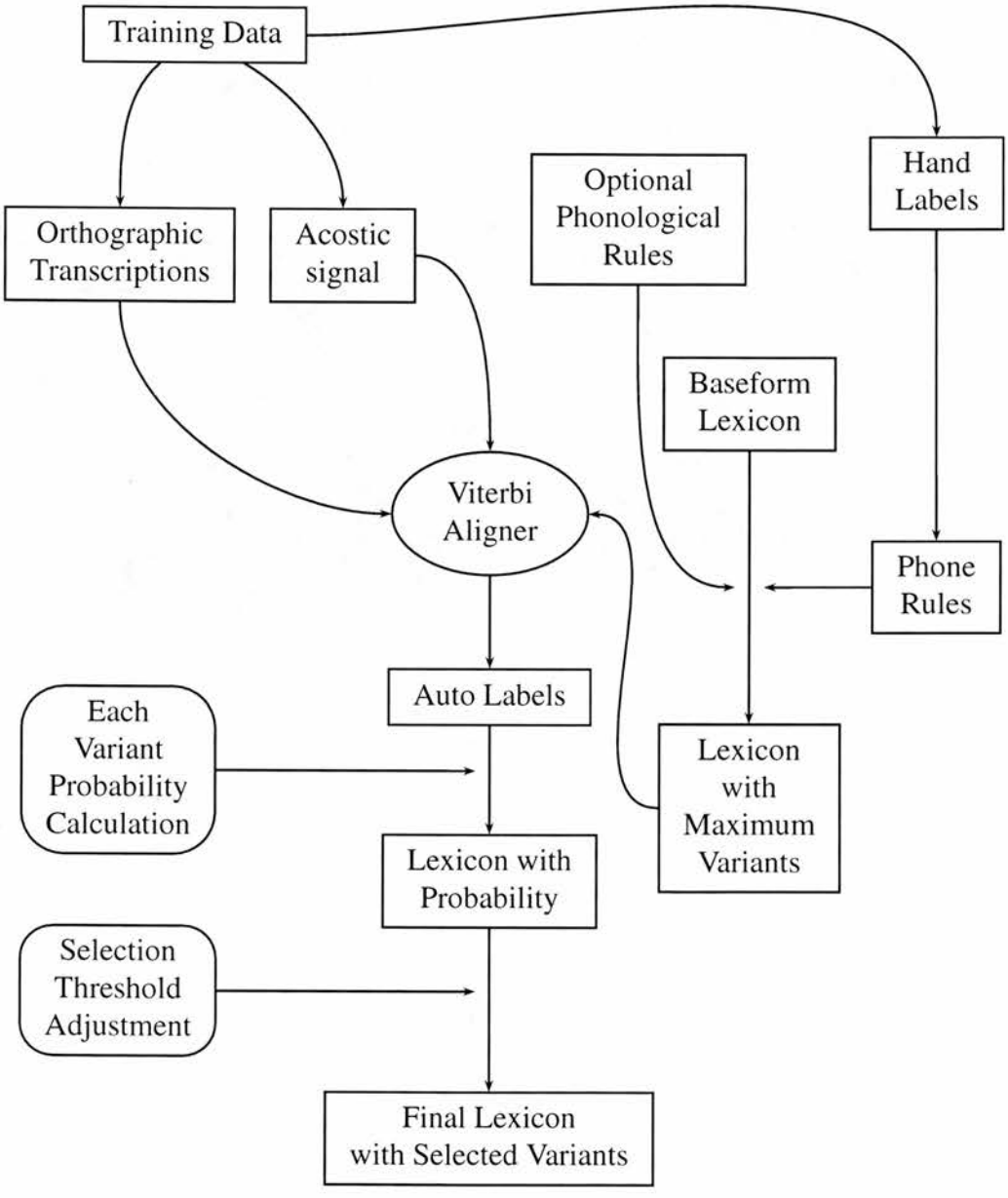


Figure 3.1: Block diagram of pronunciation variation generation.

3.9.3 Generating variations

Multiple variants of pronunciation for each item in the base lexicon can be generated in terms of various phonological and phonetic processes. The key problem is to find such rules. The rules here are different from those used in establishing base lexicon in that their application is optional. I used two ways to find optional rules.

Allophonic rules: top-down approach

First, allophonic rules described in earlier phonetic and phonological studies of Korean (Huh 1991, Lee 1996) are examined and some of them are employed after necessary modifications (top-down approach).

Phone rules: bottom-up approach

Those allophonic rules, however, do not appear to be sufficient to cover many other variations produced by other factors like speaker variability or speech style. To take this into account, I constructed another temporary version of the lexicon on the basis of phone strings created by hand labelling (bottom-up approach). Then each item in this 'hand label lexicon' is compared with the corresponding item in the base lexicon. If any difference is found between phone strings of the same lexical item in the two dictionaries, it means a new pronunciation variant is found. As long as this variant is not thought to be a trivial mistake in speech, a new phonological rule is established to account for the variant considering its context. Here, judgement is based upon both intuition and frequency of the variant. Take (3.13) for example, the phone strings of a word 'kwi-sa' are found to be different in two lexicons such as

(3.13)

Word	kwi-sa 'your company'
In base lexicon:	k wi s a
In hand label lexicon:	k wi s a (27)
	k i s a (5)

	Phonological Rule	Conversion Example	
a.	Vowel deletion	o s i p → o s p	'fifty
b.	Glide deletion	s a m wv l → s a m v l	'March'
c.	h-deletion	s e s i m h a n → s e s i m a n	'careful'
d.	Place assimilation	i m n i t a → i m m i t a	'Be+RESPECT+END'
e.	Cluster simplification	p^h u m m o k → p^h u m o k	'item'
f.	Aspiration	s i l t a → s i l t^h a	'hate+END'
g.	Tensification	h yo k wa → h yo k' wa	'effect'

Table 3.3: Phonological rules for generating pronunciation variants. Each rule can be subcategorised.

The hand label lexicon contains two pronunciation variants, one of which is the canonical pronunciation [k wi s a], and the other is a new variant [k i s a]. The numbers at the end of each item give the count of such a pronunciation being observed. Given this information, we can see that there is a variation caused by a *glide deletion rule*. Note that this rule is only optional as the canonical pronunciation also occurs quite often. The observation that the varied pronunciation occurs 5 times out of total 32 cases (15%) indicates that this variant is not just a slip of the tongue made by mistake. A number of deletion and assimilation processes are detected through this method. I will call this type of rules, 'phone rules' contrary to the existing allophonic rules.

By combining the above two types of rules together, a list of phonological rules is established as summarised in Table 3.3.

Some of the rules need to be ordered. For example, a base lexicon phone string [i m n i t a] is sometimes realised as [i m i t a]. To account for this, 'place assimilation' should apply first followed by 'cluster simplification'. If the order is reversed, there is nothing for the 'cluster simplification' rule to apply to and the intended surface form is not derived. This can be illustrated as,

(3.14)

- base lexicon item: i m n i t a
- place assimilation: i m m i t a
  cluster simplification: i m i t a (surface form - OK)
- cluster simplification: -
  place assimilation: i m m i t a (surface form - incomplete)

So far, only word internal variations are dealt with. Research has shown that across-word modelling of variation also helps decrease word error rate (Beulen *et al.* 1998). In Korean as well, some rules are influential across word boundaries. In the KAIST data, for example, a word final alveolar nasal [n] can undergo a 'place assimilation' rule and be changed into a bilabial [m] or velar [ng] nasal, due to the first phone of the next word. This type of variant is also included.

Finally, a 'variation lexicon' is established simply by joining together all the variants generated by the above methods with the canonical items in the base lexicon[3]. The number of variants for a lexical item depends on the number of phone rules imposed on it. When N rules are applicable, a 2^N variants, including the base form, are generated. This implies that longer words tend to have more variants as more than one rule may be applicable. For the KAIST data, the total number of items generated by the above phone rules is 87094 (from 2920 base lexical items).

3.9.4 Selection of variants

It is impractical and unnecessary to compose a lexicon including all the pronunciation variants generated in the above way. Doing so will result in severe inefficiency of recognition performance as the search path of the decoder will become unreasonably large.

[3]Even if some canonical items were never found in recognised speech, they were not excluded as they might appear in more controlled speech like isolated word pronunciation.

Thus, it is necessary to reduce the size of the dictionary. The reduction should be attempted in a systematic way to ensure that only less likely variants, from the point of view of recogniser, are eliminated with major variants retained.

One possible way is to use the count of each lexical item in the hand label lexicon in such a way that if a variant appears less than a certain threshold it is discarded. But this kind of treatment is not appropriate because the variants are selected based on manual annotation. First, as the number of label files is relatively small, many variants appear only once. To remove all these variants will cause loss of generality. Doing enough hand labelling to prevent this is quite costly and error-prone. Second, and more important, it is likely that hand labels may not effectively reflect the performance of the recogniser. Riley *et al.* (1998:115) show from recognition tests that "the hand-labelled data is ... not reliable enough for directly estimating pronunciation models." Strik & Cucchiarini (1998:138) also indicates "... transcriptions automatically obtained ... are more in line with the phone strings obtained later during recognition with the same ASR."

Given those insights, I calculate the frequency of each variant from the automatically generated labels of the whole training data instead of hand labels. (See page 78 for detailed description of auto labelling). Since there is a fairly good number of tokens in the training data, the first problem, shortage of data, is assumed to be taken care of. The second problem is also resolved since auto labels are products of the recogniser itself.

The 'auto label lexicon' is created in the same way as the hand label lexicon (3.13) was created. Again, the frequency of each variation in the data is also counted and specified as

(3.15)

Word	kwi-sa 'your company'
In base lexicon:	k wi s a
In auto label lexicon:	k wi s a (208)
	k i s a (127)

Note an increase of the number of word occurrence (335) in comparison to that of hand labels (32 in (3.13)), which makes statistical inference more meaningful. It should also be pointed out that a recogniser more often chooses a variant which underwent the glide deletion rule (37.9%) than human labellers (15.6%). This justifies that using auto labels is necessary to determine whether a rule or its output pronunciation is helpful for speech recognition.

Based on this 'auto label lexicon', selection of variants for each word item W is made according to the procedure given below.

(3.16)

- The base form is always kept
- For each of the remaining variants V_i, the probability $P(V_i|W)$ is calculated as

$$P(V_i|W) = \frac{C(V_i)}{C(W)}$$

  where $C()$ is a counting function.
- Discard V_i if $P(V_i|W) < T$ where T is a cutoff threshold

The threshold T can be heuristically adjusted depending upon the size of lexicon and recognition accuracy. A bigger T value will result in a smaller size lexicon. I found that 0.1 is a reasonable value to reduce the size of the lexicon considerably and still retain major variants. In this way, a new version of the 'variation lexicon', which contains 12059 items (reduced from 87094 items), was constructed.

3.9.5 Evaluation

A recognition test was performed to see whether the new lexicon is useful in practice. Apart from the lexicon, exactly the same data, methods, and language models were used as in section 3.8. The word accuracy in the current experiment is 87.93 %, which is a

considerable improvement compared with the initial system with the baseform dictionary whose accuracy was 78.47%.

It needs to be mentioned that judging performance of this recogniser by comparing with other Korean recognisers is difficult. Obviously, one of difficulties of objective evaluation or comparison in Korean speech recognition is lack of databases. There are few, if any, database resources for a large vocabulary speech recognition system which are available in public. Consequently, most reported recognition experiments employ a database of the researchers' own development, which is rarely released (or maybe rarely wanted by others) even after the experiments. This difficulty in finding a quality database has made researchers spend much time on the construction of data corpora necessary for their research, which usually gives rise to longer development time and often is a poor use of researchers' time.

In the literature, I found one report of a speech recognition experiment (Choi *et al.* 1995) in which the same KAIST data is used. The recognition accuracy of their system in terms of triphone modelling was 90.7% which is better than the current system. However, a direct comparison is not available since they used a subset of the data composed only of male voices for both training and test, while the current system is gender independent. Thus, I assume that the performance of baseline model described in this chapter is fair enough to be a benchmark for later comparison with other performance.

3.10 Summary

A baseline recogniser is constructed based on context dependent phone units and HMM recognition techniques. Various methods are employed for maximising performance. For the improvement of recognition performance, the lexicon is modified to contain alternative pronunciations. Both high-level and low-level information sources are utilised for systematic selection of frequent variants. The accuracy increase indicates that the lexicon modification is valuable.

The baseline recogniser is supposed to contribute to the comparison of the systems using the proposed segmental F0 effect. In addition, the various specific techniques, by-products of the recogniser construction described in this chapter, will also be applied in other experiments.

CHAPTER 4

Structure of Segmental F0

As mentioned in the Introduction, segmental F0 effect in Korean can be described as: the vowel F0 is low after weak sounds such as lax obstruents and higher after strong sounds such as tense or aspirated obstruents.

In this chapter, I present automatic analyses of this effect in Korean. The first order of business is to verify the effect itself. Though there have been a number of phonetic experiments on this effect most of them are drawn from non-Korean data, in which the voice/voiceless contrast is relevant instead of three way contrasts of Korean. A few studies have dealt with Korean data, but their results cannot be considered conclusive due to the small amount of data usually used (see section 4.1 below). Furthermore, some results are not in agreement with each other, calling for further investigation.

After verification of the segmental F0 effect, I will attempt to use the statistical results in the automatic manner classification of Korean stops and affricates, in the second part of the chapter. The classification performance of segmental F0 will be compared with that of spectral information used in standard speech recognition tasks. Though this experiment only deals with manner identification of obstruents, the manner correctness is obtained largely in terms of analysing initial outputs of speech recognisers. Thus, the results can be regarded as a preliminary verification for further application to speech recognition.

Article	Context	LAX	TNS	ASP	Tokens
Kim 1965	stop+V+C(?)	173.91	194.17	188.68	5
Han & Weitzman 1970	stop+V (male)	144-162	178-191	185-201	?
	stop+V (female)	266-309	308-337	341-343	
Hardcastle 1973	stop+e	170-185	200-204	182-192	126
	stop+u	172-179	204-213	200-213	126
Kagaya 1974	stop+[e or i]	148	160	162	12
	affricate+[e or i]	144	157	157	4

Table 4.1: Summary of earlier experiments on segmental F0 effects in word initial position. Each line is an average of single speaker tokens. The original values, which were specified in terms of one full cycle period, are converted into frequency (Hz) for better comparison. All values are rounded to the nearest Hertz.

4.1 Review

Table 4.1 is a summary of earlier experiments on segmental F0 effects.

Kim (1965:349) seems to be the first attempt at measuring segmental F0 effects in Korean, though his interest was mainly focused on finding other cues like tensity or voicing. He measured the distance between the two glottal pulses of vowel onsets on an oscilloscope display. The average durations of one full cycle he reported are 5.75 msec after lax, 5.15 msec after tense, and 5.30 after aspirated stops, which are converted into corresponding F0 values in Table 4.1. It is interesting that in his measurements tense stops had higher F0 than aspirated stops, while the reverse is the case in some later experiments. Han & Weitzman (1967, 1970) also report word initial F0 values of isolated words. The F0 following aspirated or tense stops is claimed to be about one and one-fourth times greater than after lax stops. There was considerable overlap between after-aspirated and after-tense F0 values. The voices of two informants were analysed but the number of averaged tokens for measurements is unknown. Hardcastle (1973), like Kim, measured periods of glottal pulses of two vowels ([e] and [u]) after consonants of each type in isolated words and obtained similar results to Kim's experiments in that tense stops give rise

to higher F0 than aspirated stops. While most reports are on the effect of stop consonants, Kagaya (1974) measured the effect of affricate consonants, too. As expected, the F0 distribution after affricates of each type turned out to be similar to that after stops. These experiments have in common that a relatively small number of tokens, usually spoken by a single speaker, were averaged. Also, visual inspection and manual calculation were used to measure F0 values.

All the results in the table agree that F0 after lax stops is distinctively low compared with F0 after other consonant types, but there is an inconsistency in the order of after-tense and after-aspirated F0. The reason for this is not clear at present but the small number of tokens and measurement error in some experiments may have caused such inconsistency.

More recently, Jun (1996) confirmed the effects in a cross linguistic experiment. She also found that the segmental F0 effect is more distinct in Korean than in European languages like English or French. Her findings regard the magnitude of F0 and the duration of the effect. First, the size of segmental F0 effect was considerably larger in Korean than in English or French. Second, the duration of the effect is longer in Korean (at least 100 msec) than the other two languages (40-60 msec). In her experiment, however, the classification of consonants was different from the previous studies reviewed above. She used a two way distinction (ie., strong vs. weak) instead of the typical three way distinction of Korean consonants[1]. Aspirated and tense stops were collapsed together in a single class, 'strong', together with fricatives, while lax stops were included in 'weak' class along with the sonorants ([m] and [l]).

4.2 Data Creation

I collected a data set composed of words with relevant stop and affricate sounds. The material to be recorded was 216 two-syllable isolated words, selected from a Korean dictionary, most of which include one of the 12 obstruents in question in both first and

[1]Probably, this is for the convenience of a direct comparison with English and French in which only two way voicing distinction is applicable.

second syllable onset position. Unlike most phonetic experiments, the vowel type of each syllable was not fixed and vowels of various different height were as evenly distributed as possible. This is helpful to find the effect of vowel intrinsic F0.

All the tokens were spoken within a fixed carrier sentence in the form of:

(4.1) *i-keos-eun* _____ *cheo-leom* *po-in-ta*
this+TOP like look
'This looks like _____'

This uniform design is for the purpose of keeping macroprosodic effects to a minimum or at least constant. Though, as already mentioned, underlying prosodic structure will not totally obliterate the segmental effects, its non-uniform intervention will undermine reliability of the statistical estimation of the effects.

Four male speakers participated in recording: JSH, KHK, TSS, and WHY. All four subjects were brought up and educated in the area where Seoul Korean is spoken, and are currently postgraduate students at the University of Edinburgh. Only one of the speakers, WHY, knew the purpose of the experiment but I assume the influence of his previous knowledge is not large enough to alter the results to any considerable degree. Sentences were recorded in randomised order. 16 bit digitisation was conducted with 16 KHz sampling rate and a total 4320 tokens (216 words x 4 speakers x 5 iterations) were obtained.

4.2.1 Automatic annotation

To analyse segmental effects in speech data, phone boundaries must be marked. An auto-labelling scheme was introduced for this purpose. Error free hand labels might provide a more accurate description but the automatic method has its advantages. First, it is fast. Once an acceptable auto labeller has been built, a large amount of data can be annotated in a relatively short time. For example, Schiel *et al.* (1998) points out that more than two hours were spent in hand labelling of a 10 sec spontaneous utterance. Second, the pattern of errors is consistent. With hand labelling, the pattern of errors tends not to be

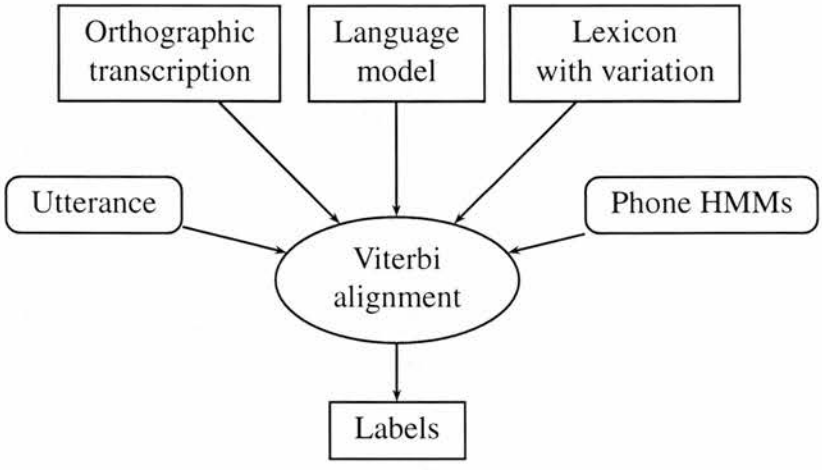


Figure 4.1: Automatic labelling procedure.

uniform and checking is difficult. On the contrary, errors caused by autolabelling, if any, are easier to detect and can usually be corrected, again, automatically.

In fact, a phone-model speech recogniser can be directly used as an autolabeller without any major modification. In this case, the recognition target is the phone string rather than the word string. It is necessary to provide

(4.2) Phone sequence language model
Orthographic transcription
Word-to-phone string lexicon

Then, alignment is forced using the *viterbi* algorithm to generate autolabels with temporal information. The procedure of auto-labelling is briefly illustrated in Figure 4.1. Providing the variation lexicon, described on page 65, instead of just baseforms, is quite useful for obtaining accurate surface phonetic labels, with various phonological rules taken into account. For instance, in case a phone in a word is optionally deletable due

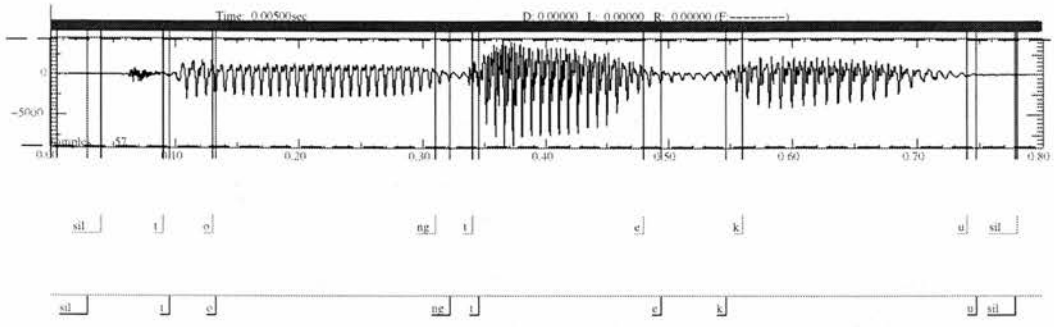


Figure 4.2: Example of auto-label: The spoken word is *tongteku* [toŋteku] ('east of Taegu city') marked as [sil t o ng t e k u]; the bottom line marks are labels created by hand and the upper marks are auto-labels.

to the speaker's speech rate we specify both phone strings, with and without the phone in question, and let the recogniser pick out the more suitable one for the given utterance token on the basis of the acoustic evidence.

I used the baseline recogniser described in chapter 3 as an autolabeller. As a phone recogniser, its phone string accuracy over unseen data was found to be 68.29%, with a bigram language model. This figure is cited just to give a general idea of the recogniser's quality. Recognition accuracy is not directly relevant to the current task, since the phone string is supplied, and only the boundaries need to be identified. All 4320 words were autolabelled in this way.

Figure 4.2 is an example of auto label output compared with the corresponding hand label. Correcting misplaced phone boundaries in the automatic segmentation might have been a way of getting sharper results in the phonetic experiments of section 4.3, but no hand correction was attempted so that the ASR system will be trained with the same type of errors that it will be tested with.

4.2.2 F0 extraction and normalisation

For F0 estimation, I used the *get_f0* pitch tracking program of *ESPS* (Entropic 1998) which is based on *normalised cross correlation* and *dynamic programming* as described in Talkin (1995). Though the performance of this detection algorithm is known to be quite robust it still frequently fails to detect some important regions around the vowel onset position where crucial information for segmental perturbation is contained. Admittedly, this problem is not tackled in this study, which implies that a better F0 estimation algorithm in the future will further improve the result of this research.

The pitch range of the four speakers cannot be assumed to be the same. Moreover, within-speaker range may not be consistent when an utterance is pronounced several times. Such variation will affect the quality of statistical inferences. Briefly, all the F0 values are redistributed based on a fixed mean value 130 Hz, which is found to be the average vowel F0 of all male speakers[2]. More detailed description and a formula for speaker normalisation are given in section 6.2.2 on page 123.

4.2.3 Measurement

A program was written to perform cumulative statistical calculation of context dependent F0 values. Given a context (eg., high vowels in word initial syllable with lax stops and affricates at the onset position), the program computes statistical parameters like *mean*, *standard deviation*, and *number of tokens*.

To represent the F0 of a vowel, the overall mean of F0 values of all the non-zero frames of the vowel period was calculated. When a positive F0 value begins to appear before the boundary between the vowel and the preceding consonant, as is often the case, the F0 values at ending frame(s) of the consonant are also counted in calculation (Figure 4.3). This adjustment is based on the general agreement that the segmental F0 effect is most

[2]For the z-score normalisation, as in the present experiment, to be effective, normality in distribution of data needs to be presupposed. My previous study (Jang 2000) confirms that segmental F0 values are normally distributed at least in Korean.

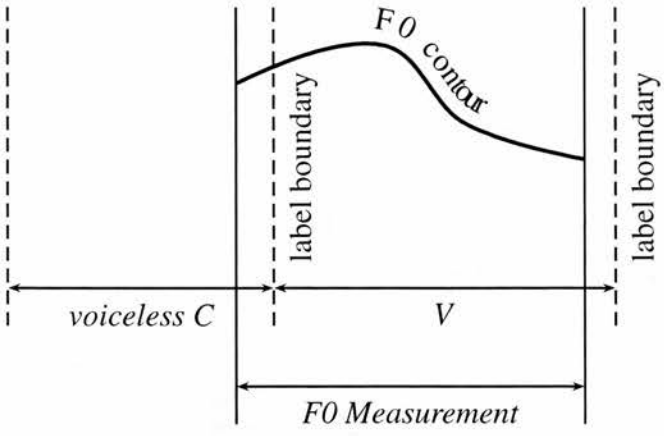


Figure 4.3: F0 measurement range.

effect at the onset of vowels. On the other hand, if the beginning frame(s) of vowels do not have positive F0 values, they are not taken into account.

Whenever a judgement of the statistical significance is necessary, an *Analysis of Variance (ANOVA) Test* is performed. As it simply determines whether any variable is statistically different from any other variable without concretely specifying the relevant variables, a supplementary test, called *Tukey's multiple comparisons* procedure (Devoer & Farnum 1999:395), is subsequently performed, which takes into account all possible pairwise comparisons. Both tests are performed at the significance level of $\alpha = 0.01$ (or 99 % confidence level), unless specified otherwise.

4.3 Extent of the Segmental F0 Effect

Two aspects make the current analysis different from earlier phonetic studies. First of all, a relatively large database is used to verify the F0 characteristics. One of the elements that make statistical modelling more reliable is a large number of speech tokens. Another advantage of using a large database is the possibility of drawing results from various

contexts for comparison. Given enough data, this can be done relatively easily saving time and effort which would be spent on careful experimental designs.

Secondly, automatic rather than manual methods of segmentation, annotation and measurement are adopted. This is inevitable for the ASR application, as when it comes to the test phase of recognising unseen data, an on-the-fly estimation of parameters must be performed automatically without any help from human intuition. I will show that automatic data processing and analysis reveals the segmental F0 effect to a statistically significant degree.

4.3.1 Overall average

Though the results of averaging all the vowel F0 values following a given stop manner type, without considering other influential variables and contexts, might be too crude to draw any decisive conclusions from, they can at least show the general tendency of the distribution. From the results shown in Table 4.2 and Figure 4.4, it is clear that the F0 of vowels following aspirated consonants is higher than the F0 of vowels following the consonants of the other classes. Likewise, F0 after tense consonants is higher than after lax consonants. While the tendency is consistent for all speakers, the magnitude varies depending upon the pitch range of each speaker. Figure 4.4, shows that speaker JSH has the highest pitch on the average. The speakers in this experiment were all male, and there would be more variation if female speakers were included. The considerable difference in pitch range even among male speakers suggests that further manipulation such as speaker normalisation will be essential for F0 to be exploited in speaker independent recognition systems.

4.3.2 After normalisation

Results after speaker normalisation are shown in Table 4.3 and Figure 4.5 for comparison with the results before normalisation in Table 4.2 and Figure 4.4. Figure 4.5 particularly shows how normalisation has been effective in minimising, if not eliminating, the speaker

Speaker	LAX	TNS	ASP
JSH	151.86	169.04	177.18
	19.47	11.34	12.49
	837	521	575
KHK	113.55	125.94	139.05
	18.31	19.63	21.43
	818	502	511
TSS	112.36	123.36	129.31
	11.61	11.25	12.14
	825	519	559
WHY	115.29	130.90	151.03
	19.92	25.34	26.40
	812	516	548

Table 4.2: speaker. Each row for each speaker represents, top to bottom, mean F0 magnitude (Hz), standard deviation, and the number of tokens. For each speaker, values of any pair of three classes are significant different ($p < 0.001$).

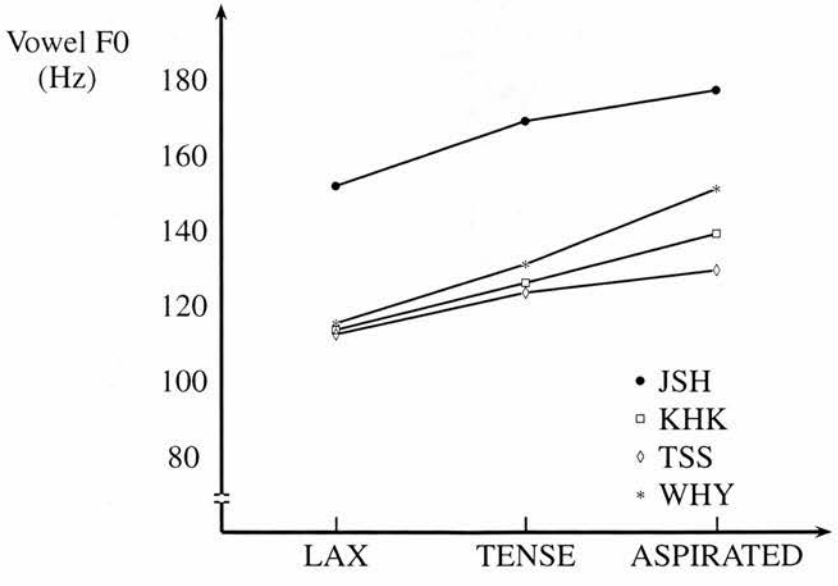


Figure 4.4: Comparison of F0 distribution of 4 speakers.

Speaker	LAX	TNS	ASP
JSH	120.86	139.61	148.26
	21.65	12.71	14.51
	837	521	574
KHK	123.04	135.32	147.92
	17.61	18.90	20.75
	823	499	512
TSS	123.51	133.88	139.14
	10.78	9.98	10.65
	823	518	559
WHY	120.11	131.70	147.40
	15.36	18.97	19.55
	815	517	544

Table 4.3: Speaker normalised F0 magnitude for each speaker. Each row for each speaker represents, top to bottom, mean F0 magnitude (Hz), standard deviation, and the number of tokens.

differences. I will, for the rest of this experiment, report only values based on normalised F0.

4.3.3 Syllable position factor

The size of the segmental F0 effect appears to differ depending upon the position of the syllable in a word. Measurements demonstrated in Table 4.4 reveal that the effect is quite distinct in the first syllable while it wanes considerably in the second syllable. This result is consistent with Jun's (1996) finding that the effect was only evident at the Accentual Phrase initial position.

In a phonetic experiment, Lee (1998) finds that F0 in later syllables depends on the initial consonant of the first syllable. This accounts for why the segmental F0 effect lose its status in word non-initial syllables. As the segmental F0 effect at the AP initial position is strong enough to persist in later syllables, F0 in the following syllable is affected

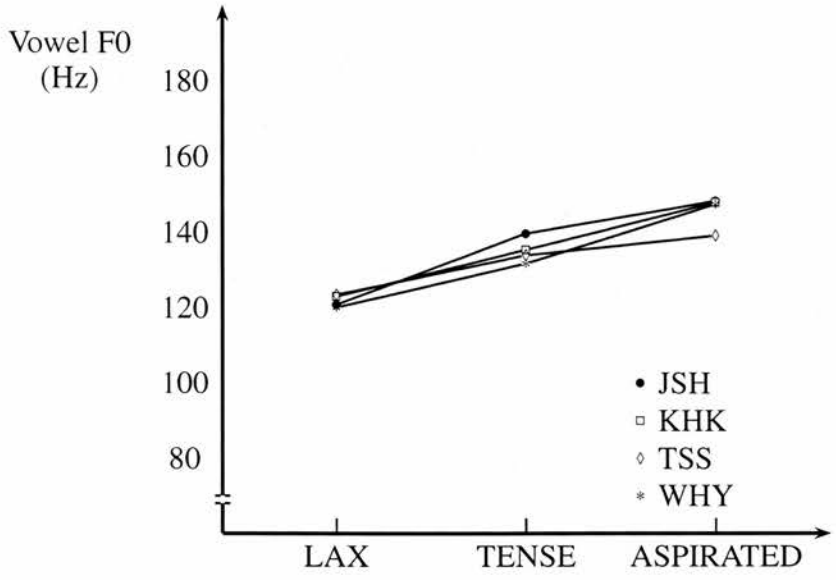


Figure 4.5: Comparison of F0 distribution of 4 speakers after normalisation.

Context	After LAX (i)	After TNS (ii)	After ASP (iii)
(a) First Syllable	112.50	146.62	157.31
	10.06	15.46	13.65
	1770	618	970
(b) Second Syllable	130.82	131.48	139.36
	18.91	15.75	17.17
	195	1439	1207

Table 4.4: Segmental F0 depending on syllable positions.

not only by its own segmental structure[3] but by the segmental structure of the preceding syllable(s). To confirm this observation, I measured the F0 of second syllables depending upon the segmental structure of the first syllable, the result of which is in Table 4.5.

In brief, F0 influence of the first syllable is quite strong so that it persists in the next syllable, preempting the second syllable's own prosodic structure. The F0 of a second syllable which has an aspirated onset consonant but is preceded by a syllable with a

[3]The macroprosodic effect is not considered here.

Context		Mean	St. Dv.	Num
1st syl	2nd syl			
LAX	LAX	115.94	7.31	68
TNS		133.78	13.41	33
ASP		149.87	15.17	49
LAX	TNS	123.43	11.93	195
TNS		148.48	15.05	173
ASP		151.18	13.33	43
LAX	ASP	126.90	10.16	292
TNS		144.72	10.97	23
ASP		153.37	13.50	248

Table 4.5: Across syllable influence of F0.

lax onset consonant, tends to be much lower than for a syllable which has a lax onset consonant but is preceded by a syllable with an aspirated consonant.

This suggests that the segmental effect is consistently useful only in the AP initial position. As an AP is composed in many cases of just one prosodic word, I assume that segmental F0 is effective in prosodic word initial position. Consequently, the exploitation of the F0 effect in chapter 6 will also be based on prosodic words.

4.3.4 Consonantal place specific distribution

The distribution of vowel F0 after each consonantal place is investigated and summarised in Table 4.6. It is shown that the way vowel F0 is affected by the manner of the preceding consonant is consistent for all places of articulation, though a little difference in magnitude is observed. It is noteworthy that the segmental F0 effect is more salient when consonants are affricates than stops.

Table 4.7 shows that half (9 pairs) of possible 18 pairs of the same class phones are significantly different at the confidence level $\alpha = 0.05$, or 28% (5 pairs) at the level $\alpha = 0.01$.

Place	Previous Phone	Mean	St. Dv.	Num
Labial	p	111.87	9.56	479
	p'	145.20	14.80	159
	p^h	157.33	12.98	226
Dental	t	112.37	10.18	308
	t'	143.40	14.53	199
	t^h	154.12	13.63	286
Velar	k	113.59	10.61	735
	k'	150.64	14.06	160
	k^h	155.70	14.19	132
Palatal (Affricate)	c	110.63	8.68	248
	c'	148.87	18.37	100
	c^h	160.73	13.08	326

Table 4.6: Consonantal place specific distribution at word initial syllable.

Phone Class	Compared pair	$\alpha = 0.05$	$\alpha = 0.01$
LAX	c - k	√	√
	k - p	√	
TNS	k' - p'	√	√
	k' - t'	√	√
	c' - t'	√	
ASP	c^h - k^h	√	√
	c^h - t^h	√	√
	c^h - p^h	√	
	p^h - t^h	√	

Table 4.7: Pairs of statistically significant difference. Significance test was in terms of the single factor ANOVA test followed by Tukey's pairwise comparison.

However, it is obvious that inter-place difference is a lot less than inter-manner difference as all the pairs with manner difference are found to be highly significant ($p < 0.001$).

4.3.5 *Vowel intrinsic F0 effect*

As mentioned in section 2.6.2, one of the F0 factors to be considered is the change of F0 value due to the height of the vowel or IF0. There is a tendency for high vowels, such as [i], [u] to have higher F0 than low vowels such as [a].

A number of phonetic experiments have verified this effect, but reported size of the effect varies. For example, Peterson & Barney (1952) found the difference between F0 of high vowels and low vowels to be as much as 25 Hz, on the average. Similarly, Lehiste & Peterson (1961) measured isolated word utterances of five speakers and found that there is an approximately 18 Hz difference between vowel [u] and [ɔ]. In House & Fairbanks (1953) only a 9 Hz difference was reported. Silverman (1987) and Petersen (1986) confirm this effect by showing that listeners, of English and Hindi respectively, also adjust for intrinsic F0 when they recover the underlying prosodic relations from the acoustic signal. For example, when a low vowel and a high vowel is produced at same F0, listeners tend to think that the low vowel is spoken with higher F0. It implies that the IF0 of low vowels are supposed to be lower than high vowels. The vowel IF0 effect has been confirmed in a large number of languages by phonetic experiments. Whalen & Levitt (1995) reviews the relevant literature on IF0 for 31 languages and compares the results in various ways. Based on such observation, they conclude that the IF0 is a universal effect, which most (virtually all) investigated languages share.

I have not found many studies on IF0 of Korean vowels. In Han & Weitzman (1967), an effect can be observed but too large a variation range between speakers makes it inappropriate to estimate the average size of the difference between high and low vowels. In fact, the main purpose of that article was not to study IF0 but segmental F0 effects.

Lehiste & Peterson (1961:424) say: "[in English] the fundamental frequency of a vowel with high intrinsic fundamental frequency occurring in a word beginning with a consonant that has a lowering influence may overlap that of a vowel with a low intrinsic fundamental frequency preceded buy a consonant that has a raising influence." If this

Syllable Position	Preceding Consonant	High V	Mid V	Low V	Low-High Significance
First Syllable (AP Initial)	Lax	116.65	112.26	105.36	$t = 18.14$ $p < .001$
		10.11	9.28	7.55	
		615	887	348	
	Tense	153.61	147.72	138.37	$t = 8.92$ $p < .001$
		14.12	14.29	15.17	
		139	319	160	
	Aspirated	160.50	157.62	155.51	$t = 3.66$ $p < .001$
		15.24	13.42	12.98	
		142	483	350	
Second Syllable (AP Non-Initial)	Lax	126.34	120.12	116.75	$t = 10.37$ $p < .001$
		18.68	16.62	17.74	
		1192	1390	590	
	Tense	136.98	137.82	129.87	$t = 6.79$ $p < .001$
		17.75	9.55	9.68	
		801	878	379	
	Aspirated	149.49	146.26	147.39	$t = 2.01$ $p = 0.045$
		17.93	18.51	17.20	
		524	1055	609	

Table 4.8: Vowel Intrinsic F0 Effect. Three rows of each context represent, top to bottom, mean F0 (Hz), standard deviation, and number of tokens. The two-tailed $t - test$ is used for significant test.

phenomenon of cancelling-out each other's effect occurs in Korean as well, the use of segmental F0 may not be available for automatic recognition without separating it out from vowel effect, which is extremely difficult. The experimental result in Table 4.8 bears on this issue.

As shown in the table, the IF0 effect is consistently observed in the current data, regardless of context. Except when vowels are preceded by an aspirated in AP non-initial syllable ($p = .045$), the F0 of high vowels are significantly different (when $\alpha = 0.01$) from that of low vowels ($p < .001$). Most importantly, however, the IF0 effect is not strong enough to preempt the segmental F0 effect in any context. That is to say, the F0

magnitude of a 'lax C + high V' syllable never exceeds the F0 of an 'aspirated C + lax V' or 'tense C + lax V' syllable. Another noteworthy result is that the effect is least prominent after an aspirated sound. This will be further discussed in section 6.3.6 on page 130.

The conclusion is that the F0 raising effect of the preceding consonant tends to overwhelm the IF0. This is a positive indication for the use of segmental F0 in consonant identification. The IF0 effect in connected speech data will be discussed in chapter 6.

4.3.6 Following consonant factor

Most of the literature is concerned with how vowel F0 is influenced by the preceding, not following consonant, assuming, as Mohr (1971:71) indicates, that the influence of a consonant on the fundamental frequency of a vowel is progressive rather than regressive. For example, Lehiste & Peterson (1961) states that, in English, the voiceless/voiced contrast of the final consonant has no significant influence on the F0 appearing on a preceding syllable nucleus. While discussing the reason for the segmental effect, Ohala (1978:29) also mentions that stops affect the pitch of following not preceding vowels.

On the other hand, Kohler (1982, 1985, 1986) claims, on the basis of a series of perception and production tests, that a raised F0 can occur due to the following /t/, as against /d/, as well as the preceding voiceless sounds in German and English. For example, when the F0 of later part of the /ay/ vowel in the stimulus 'widen' was manipulated to be higher, listeners were likely to respond more often as 'whiten'.

Given such disagreement on postvocalic influence, and given the fact that no previous experimental study of Korean dealing with this issue has been reported, I have looked for this phenomenon in my Korean data.

First Syllable	Second Syllable	Mean	St.Dv	Num
LAX	LAX	111.90	9.83	648
	TNS	109.87	9.24	178
	ASP	112.98	10.14	258
TNS	LAX	144.52	15.46	168
	TNS	149.07	15.15	138
	ASP	146.96	16.09	37
ASP	LAX	155.77	14.10	331
	TNS	154.35	12.64	42
	ASP	155.90	13.31	241

Table 4.9: Pre-closure vowel F0 effect. Values are F0 of vowels in the first syllable.

First, I averaged the F0 of vowels in word initial syllables depending upon the manner type of the consonants at the onset position of the next syllable. Table 4.9 shows the results of this averaging.

No striking difference of value is apparent due to the type of the second syllable and significance tests for each class values also confirm that there is no statistical significance at the 99% confidence level between any pair with the same first syllable class. ($F(2,1081) = 5.31$ for the initial LAX, $F(2,340) = 3.28$ for initial TNS, and for $F(2,611) = 0.23$ for the initial ASP, $p > 0.01$ for all cases).

Thus, the conclusion can be drawn that only pre-vocalic consonants substantially influence F0 of Korean vowels.

4.3.7 Summary

Various properties of segmental F0 have been verified or established. First, the difference of vowel F0 depending upon three types of Korean obstruents can be effectively extracted from automatic measurements. It is consistently revealed that F0 is high after aspirated sounds, low after lax sounds, and in between after tense sounds. Second,

speaker normalisation is effective in that it makes it possible to analyse the data independent of speaker range without adversely affecting the statistical distribution of segmental F0 effects. Third, the segmental F0 effect is most prominent in word initial position. Fourth, there is a vowel intrinsic F0 effect in Korean but it is not strong enough to obliterate or cancel out segmental F0. Fifth, the segmental F0 effect is mainly progressive that the influence of the following segment is negligible.

4.4 Form of F0 Contours

In this section, I analyse the local shape of F0 contours after obstruent consonants. This is to see whether there is any other exploitable feature in addition to magnitudes of the original F0 values. It is sometimes the case in speech recognition work that dynamic features like slope and acceleration of acoustic features (eg., cepstral coefficients) improve the recogniser's capacity considerably thanks to the fact that dynamic features are less vulnerable to many kinds of variability (Dumouchel & O'Shaughnessy 1993:2196).

It could be very helpful if characteristics of F0 movement could be found to depend on consonantal class. Two aspects of contours, *shape* and *slope*, will be examined. But brief reviews on the controversy related to this issue will be given first.

4.4.1 Review

There is disagreement, in the literature, on what F0 contours look like when it is influenced by the preceding consonant. On one hand, investigators report that there is a so-called 'rise-fall dichotomy'. That is, F0 contours keep falling from the vowel onset after voiceless consonants, whereas they rise for a period before falling after voiced ones (Lehiste & Peterson 1961, Haggard *et al.* 1970, Gandour 1974, Hombert 1978, Lea 1980).

On the other hand, others argue that there is no such dichotomy of direction and the post-stop contours basically move downward regardless of the consonantal class (Kohler 1982,

Ohde 1984, Silverman 1986, 1987). They state that the apparent rise-fall dichotomy can be attributed to failure to control other prosodic factors. Indirect support for no-dichotomy is found in other experiments which were not intended to investigate the issue in question. In the experiments of Umeda (1981) and Löfqvist (1975), it is shown that F0 after voiceless stops can rise for a short duration depending on the location of prosodic factors like lexical or sentence stress.

There are similar contradictory reports of experiments with Korean data. For example, Han & Weitzman (1970) state that F0 after lax stops begins at a relatively low level and then rises to a relative peak in 50-100 msec while F0 after tense or aspirated stops begins at a high value and stays constant or begins to fall within the same time span (page 116).

On the contrary, Kim (1968) (cited in Mohr 1971) found that F0 contours after all three classes were always falling, though the slope is steepest after the aspirated stops and least steep after the lax stops. Jun (1996) also demonstrates that F0 values are consistently falling after all three types of consonant.

4.4.2 Shape

Tilt analysis

To analyse the shape of the F0 contours I used the *Tilt intonation model* (Taylor & Black 1994, Taylor 2000). An advantage of the Tilt model is that its parameters can be extracted automatically.

As Figure 4.7 illustrates, the Tilt model characterises the relative shape of an F0 contour by a single number which ranges from -1 to 1. As the shape of a contour gets closer to a pure rise the tilt value gets near 1, and pure fall near -1. The tilt value is obtained from the four parameters, *Rise Duration*, *Fall Duration*, *Rise Amplitude*, and *Fall Amplitude*, which are schematised in Figure 4.6 by

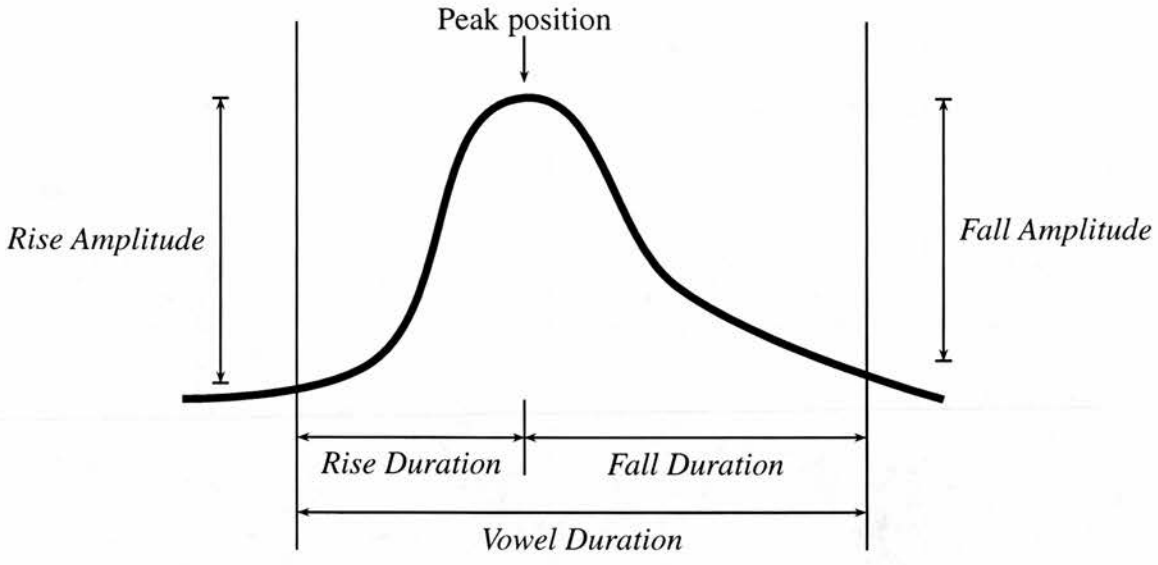


Figure 4.6: Elements of tilt parameter calculation.

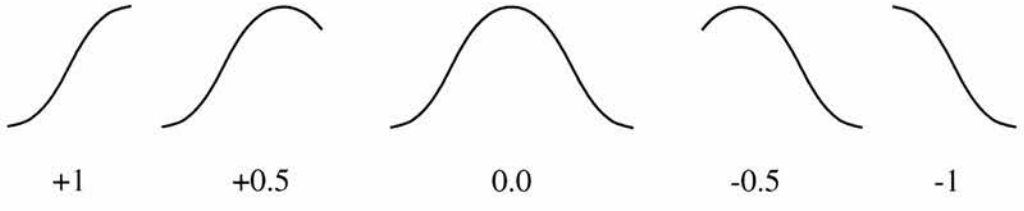


Figure 4.7: Examples of Tilt values and corresponding shape of F0 contours.

	Mean	StDv	Num	% of Negative value
LAX	-0.74	0.56	1749	88.56
TNS	-0.68	0.58	618	88.19
ASP	-0.68	0.49	944	90.25

Table 4.10: Tilt analysis: shape of the contour.

(4.3)

$$tilt = \frac{1}{2}\left(\frac{\mid A_{rise} \mid - \mid A_{fall} \mid}{\mid A_{rise} \mid + \mid A_{fall} \mid} + \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}}\right)$$

where A denotes $Amplitude$ and D denotes $Duration$.

Experiment

The same database as in the previous section is used in this experiment. Tilt parameters of F0 contours of each vowel after stops and affricates in word initial syllables are calculated. Again, the statistical significance was calculated in terms of two-tailed single factor ANOVA test and subsequently Tukey's pairwise judgement.

The results summarised in Table 4.10 show that the F0 contours after consonants of all three types have negative tilt values. This means that the shape of contours is consistently falling regardless of the type of preceding consonant. The rate of negative value is a little higher when preceded by aspirated sounds (90.25%) but the extent of difference from other classes is small. The significance test of values in each class reveals that there is no important distinction between any pair of two classes ($F(2,3308) = 4.83$, $p > 0.01$). Consequently, we can conclude that it is inappropriate to use the shape of the contour in differentiating consonant types.

4.4.3 Slope

Though all three types of obstruents turned out to give fall-like F0 contours in the following vowel, the rate of fall might be different. To see if there was a difference, I calculated the slope of each contour using linear regression. As many contours are close to pure fall, this seems a reasonable approach.

The slopes were found to be -3.27 after lax, -2.29 after tense, and -4.14 after aspirated sounds. The difference for each pair is highly significant ($F(2,3308) = 48.23$, $p < 0.001$). The steeper slope of post-aspirated F0 might be expected considering the segmental F0 effect (ie., high F0 after aspirated sounds). However, the least steep slope after tense sounds is then initially surprising since the tense sounds also give rise to higher F0 than lax sounds. The explanation for this can be sought in the difference in vowel duration which affects the slope of the contour. That is, the duration of vowels after tense sounds (78.77 msec) were found to be considerably longer than after lax (69.61 msec) or aspirated sounds (59.35 msec) [4].

The result suggests that use of slope values of F0 may be helpful as additional cues for consonant manner classification. In addition, it is also likely that vowel duration, apart from the duration of the consonant itself, can be a supplementary cue for tense sound identification, though this will not be further investigated in the current study.

4.4.4 Summary

Two aspects of segmental F0 contours are investigated. First, the shape of the contour is found to be close to a pure fall regardless of the class of the preceding consonant. So it seems that at least in Korean, there is no rise-fall dichotomy of segmentally influenced F0. Second, the difference in the slope of contours for the three manner classes is statistically significant, but it remains to be seen whether this difference is enough to be exploited in a speech recogniser.

[4] Only syllables without coda (ie.,CV) are calculated for consistency.

4.5 Automatic Manner Classification

It is not always the case that a statistically significant phonetic effect can be exploited in a speech recogniser. In this section, an experiment is performed to see if the statistical distribution of F0 shown in the earlier part of this chapter is useful in the automatic manner classification of stops and affricates at word initial position. It should be noted that as the purpose of this experiment is merely to confirm the capability of vowel F0 in identifying the class of the previous consonant, comparison of results will be focused not on the phone recognition accuracy but on the manner classification correctness.

4.5.1 Data

The same database in the previous section is used in this experiment. It is divided into two subsets, one for training and the other for test. Each subset consists of 3832 and 488 isolated word tokens respectively. All the test tokens are randomly selected. The numbers of relevant stops and affricates at the word initial position are as follow.

	LAX	TNS	ASP
Training	1552	536	858
Test	218	82	112
Total	1770	618	970

4.5.2 Methods of classification

Two sets of phone HMMs are established to perform two separate versions of automatic classification. The results of the two classification tests will be compared.

Baseline classification

The first phone HMM set is used for generating baseline results without any consideration of the information on post-obstruent vowel F0. Thus, it can be said that such standard spectral information as MFCCs, energy, and their derivatives are used for manner

classification. The recognition results obtained from those ordinary models are for the comparison with the results from the segmental F0 differentiation model. The standard phone recogniser is constructed using the methods described in section 2.7 on page 32. In the test phase, normal phone recognition is performed producing a phone string for each test token. A simple phone-bigram grammar model is generated and used [5]. When the manner of a recognised phone in word initial position is in accordance with the manner of the phone in the reference labels, it is regarded correct even if the there is an error in place of articulation. For example, when a phone [p^h] is identified as [t^h], or [k^h], it is regarded as a correct classification of manner.

Classification using segmental F0

The other classification test employs only F0 for manner distinction. For an initial step, however, a phone recogniser is also established. The HMM's used here are the same as those of the baseline system except that the models of stops and affricates do not have any manner distinction. In other words, these phones are trained and identified only as C, P, T, or K, representing voiceless affricate, bilabial stop, alveolar stop, and velar stop. In parallel, an F0 file for each test token is generated and then normalisation is performed. The normalised post-consonantal vowel F0 for each vowel, as identified by the recogniser, is used to obtain likelihoods for each of manners for the consonant preceding the vowel by the standard Bayesian formula

$$P(Class|F0) \simeq \log P(F0|Class) + \alpha \log P(Class)$$

where the weight α is heuristically adjusted in order not to have the likelihood score dominated by prior probability of each class (ie., $P(Class)$). Finally, the decision of the manner of the consonant is made on the basis of this likelihood. This procedure is

[5] Different types of language models such as an artificial finite state grammar or a phone pair grammar, did not result in any considerable difference in the interpretation of compared classification results.

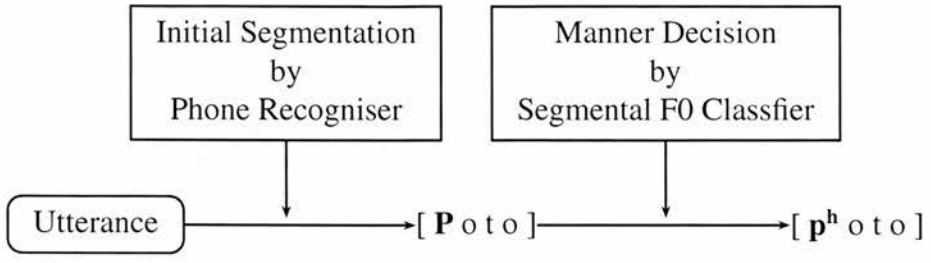


Figure 4.8: Procedure of Phone recognition by manner classification using F0.

Type of distinction	Normal recogniser	Segmental F0
LAX/TNS/ASP	86.20	78.23
LAX/ASP	84.72	99.49
LAX/TNS	97.15	91.02
ASP/TNS	100	65.52
Weak/Strong	87.57	95.08

Table 4.11: Comparison of Manner Classification Correctness (%).

schematised in Figure 4.8 where the phone string of an example word *po-to* 'grape' is recognised.

4.5.3 Results

Table 4.11 compares results produced by the two methods of classification. Five comparisons are presented in this table. First, the three way distinction considers all three manners of articulation for calculating correctness. For example, when a sound's manner is identified as either of the other two wrong manners, it is regarded as incorrect. The next three rows are pairwise distinctions, ignoring one category for each case. Finally, in the Weak-Strong distinction, aspirated and tense sounds are collapsed into a single category "Strong". Thus, a tense sound classified as aspirated or an aspirated sound classified as tense is not regarded as incorrectly classified.

It is found that the segmental F0 cues are very useful for the lax/aspirated distinction (99.49% correctness). Although the spectral information used in the normal speech recogniser is better in making the overall three way distinction, the relatively poor performance of the classification using segmental F0 is obviously attributable to failure in identifying tense sounds or misidentifying aspirated sounds as tense. In particular, the aspirated/tense confusion is much worse when using the segmental F0 information compared with spectral information (100% correctness). This is confirmed by the results of the two way (weak/strong) distinction where tense sounds are collapsed with aspirated sounds to make a single class. As shown before (page 37), the appearance of tense sounds in Korean texts or utterances are relatively infrequent, although the current data is artificially designed to contain many tokens of all three classes. This implies that to enhance speech recognition performance on natural data, the lax/aspirated or weak/strong distinction should be practically more important than tense sound identification. Therefore, the role of segmental F0, which is more useful for these lax/aspirated and weak/strong distinction, is expected to be useful in speech recognition.

Nevertheless, it is undeniable that spectral cues as well as segmental F0 cues are also quite helpful in manner identification of consonants manners. Consequently, when both spectral cues and F0 cues are combined together to compensate for each other's drawbacks, best performance will be achieved in speech recognition of Korean. The experiment in chapter 6 will be designed in this way.

4.6 Conclusion

Automatic statistical methods of F0 analysis are found to be useful in capturing the pattern of vowel F0 fluctuation influenced by the manner of articulation of the preceding consonant in Korean. It turns out that F0 of a vowel following an aspirated stop or affricate is significantly higher than when it is following a corresponding lax counterpart. The effect is particularly boosted at the initial syllable of each word token while it wanes at non-initial positions.

As for the automatic classification of lax/aspirated stops and affricates in word-initial syllables, the segmental F0 model only using F0 information of the following vowel shows better performance than the baseline model created from the standard HMM technique using various other conventional parameters. This leads to an expectation that through combination with other standard parameters, segmental F0 information can improve the quality of ASR systems.

CHAPTER 5

Speech Recognition with Demisyllable Units

5.1 Introduction

As we have seen that the segmental F0 effect is detected by statistical inference using automatic measurements, the next task is to find an appropriate method by which this F0 effect can be implemented into an automatic recogniser. One imaginable way is to further manipulate the standard phone models using the F0 feature. But as King (1998:3) points out, using F0 in the same way as cepstral coefficients is not appropriate when phone-like segments are used as recognition units, because the domain of F0 variation is not segmental but suprasegmental. Thus, further complication is inevitable to take into account the F0 information of neighbouring sounds.

In the work described in this chapter, I use demisyllables as the basic units for speech recognition. A direct advantage of demisyllable units is that the segmental F0 can be included in the unit itself. However, this cannot by itself justify abandoning the standard phone units that are most commonly used in automatic recognition. In other words, it is necessary to show that the recognisers using demisyllable units can be at least as good as recognisers using phone units. Thus, I will describe a recognition experiment, in this chapter, which will show that, in Korean, demisyllables are adequate recognition units

even before considering the segmental F0 effects. The experiment is also valuable in the sense that no Korean speech recognition experiments have so far reported serious and successful introduction of demisyllable units.

As the methods of building recognition system are in many aspects similar to those methods used in the phone baseline model, the focus of this chapter will be on the procedure for generating demisyllable models from existing phone models.

5.2 Definition and Origin

A *demisyllable* can be defined as 'the interval from the beginning of a syllable to the centre of the nucleus or from the centre of the nucleus to the end of the syllable'. The term *half syllable* is often used in the similar sense but a demisyllable need not be exactly half the length of a syllable.

Fujimura (1976) was the first to suggest the demisyllable as an important linguistic unit inspired by its convenience in accounting for phone sequence constraints within syllables in English. His notion of demisyllable is preceded by an analysis of syllables. He claims that a syllable can be divided into *core* and *affix*. Syllable cores consist of a vowel nucleus and the limited number of optional onset consonants and codas which are optional. For example, an English syllable core can be expressed as $(C_i)V(C_f)$, where (C_i) stands for a consonant or a consonant cluster in initial position while (C_f) stands for a consonant or a consonant cluster in final position. An example of an affix is the final sound /s/ in the word *kicks* /kiks/. Though his justification of the affix-core division of syllable structure is mainly based on the observation of segmental sequence constraints within a syllable, which will not further discussed in this thesis (See Fujimura & Erickson (1997) for details), and does not have direct relevance to speech recognition units, that division has been conveniently utilised in the later ASR applications by others (See section 5.6 on page 107). I will also use, though not in exactly the same way, this notion in order to decrease the number of demisyllable units in section 5.7.2 below.

5.3 Syllable Units

To understand the advantages of using demisyllables as recognition units, advantages of the whole syllable unit need to be considered first as many properties of demisyllable are inherited directly from syllables.

First of all, syllables are suitable for capturing coarticulation, which is one of the major causes of degrading speech recognition performance. In recognition based on standard phone models, this difficulty can also be alleviated by modifying units into triphones, as has been done in chapter 3. But there remains a problem of appropriately reducing the inventory of triphones for efficiency. Syllables do not have such a problem. Second, syllables are more stable than phones for model training. Phonetic processes such as phone devoicing, deletion, or total assimilation to a neighbouring sound can negatively affect automatic unit alignment, which is a conventional way to train phone models without manual transcription, and thereby corrupt the quality of trained models. But a longer unit like a syllable is less likely to lose its identity, since many phonological processes only affect the segmental structure within the syllable. Third, syllables are suitable for extracting some acoustic characteristics such as suprasegmental features. This is why a number of speech recognition systems for monosyllabic tone languages (eg., Chinese) are developed on the basis of syllable like units (Hu *et al.* 1996). Furthermore, it is likely that longer units will make it easier to exploit temporal characteristics together with other acoustic cues. There is a widespread agreement that duration can be another useful cue for enhancing recognition performance. Though much of the phonetic evidence is on the temporal characteristics of individual phones, its exploitation has turned out to be difficult partly because of the short duration of phone units in general (Ganapathiraju *et al.* 1997). Some temporal cues concerned with syllable units (eg., pre-pause lengthening) will be easier to exploit thanks to the relatively long overall duration.

However, there is a disadvantage of syllable units. The inventory is generally large. In English, for example, there are approximately 10,000 syllables, compared with ap-

proximately 50 context independent phone like units (Rabiner & Juang 1993:436). In Korean, the number of syllables is a lot smaller (theoretically 3,520, according to Huh (1991:229))[1] than in English thanks to relatively simple underlying structure of syllable constituents which does not allow consonant clusters in onset or coda position. But that is still a large number of units for which to obtain sufficient training data[2].

5.4 Demisyllable Units

Demisyllables are appropriate for speech recognition units in that they share in general the above advantages of syllables, on one hand, while they alleviate the burden of obtaining a large amount of training data for syllable units, on the other. The number of English demisyllable units is substantially smaller than the number of syllable units: approximately 2000 (Owens 1993:107).

Even though the length of a demisyllable is shorter than that of its original syllable, it still includes the region of coarticulation of either the onset-nucleus transition or the nucleus-coda transition. One of the two demarcation boundaries of a demisyllable is usually located in the stationary area of the syllable nucleus. As syllable nuclei are comparatively easier to automatically detect than other syllable subconstituents like onsets or codas, segmentation of demisyllables should be easy as well. Segmentation does not have to be precise as long as the division point is in the vowel's stable period.

5.5 Demisyllable and Segmental F0

One of the motivations for using demisyllables as recognition units in this research is that they make it straightforward to incorporate F0 cues in recognition. As has been shown, segmental F0 perturbation is observed most prominently at the onset and front part of

[1]The number can vary depending on the phone inventory considered.

[2]For comparison, the number of generalised triphones used before is 648.

the following vowel. Consequently, the demisyllable is undoubtedly a suitable unit to accommodate information on consonant-vowel relationship.

5.6 Previous Studies

To the best of my knowledge, Rosenberg *et al.* (1983) is the first serious attempt to use demisyllables in a practical speech recognition system. Their introduction of demisyllables was to explore the possibility of replacing whole word units with subword units which would be capable of representing any word. They also intended to reduce the size of syllable inventories considerably. For this, they created 946 demisyllable prototypes after recording a word-level database devised to contain all target syllables. They were able to reduce the inventory of demisyllables (as well as syllables) by adopting the Fujimura's "core + affix" analysis of syllables. For example, the word *tax* was represented by demisyllables such as "TAE + AEK + S" in stead of "TAE + AEKS". The prototypes were parameterised as LPC coefficients and the *dynamic time warping* (Owens 1993:140) was used for the recognition phase. Input test words were compared with word-level templates built from demisyllable prototype. Tests with isolated word utterances showed that recognition using demisyllable models (WER 18-33%) was worse than with whole word models (WER 6-15%). Nevertheless, it was claimed that the advantages of potential extension to vocabulary independent application and of computational storage reduction were sufficient motivations for demisyllables to replace the whole word units in future work.

Ruske (1986) and Ruske & Weigel (1992) utilised demisyllables for German speech recognition. In the former, 85% correctness of words were reported with a 75 word vocabulary continuous speech data. In the latter, they reported that 96% of words and 74% of sentences were correctly recognised over a test dataset consisting of 32 sentences spoken eight times. Strictly speaking, however, these works are not genuine demisyllable based recognition, since they exploited demisyllables only as segmentation and processing units, but not as decision units. To reduce the number of units, they separated

a demisyllable into consonant cluster and vowel. For example, each consonant cluster HMM is trained in the context with all vowels, expecting that the model will become independent from the vowel identity.

Yoshida *et al.* (1989) is relevant to the present work in that the target language is Japanese which has a simple syllable structure comparable to Korean[3] and that continuous density HMMs were used for estimating output probabilities of models as in the current system. For isolated word recognition with an 1800 word vocabulary, average 97.5% correctness was reported.

More recently, Hungarian was also tested with HMM demisyllable models (Fegyó & Tatai 1999). Hungarian has in common with Korean in that its word formation is mainly in terms of morpheme agglutination. A number of derivational and inflectional morphemes can be sequentially attached to a single stem to make a word in both languages. 100 words spoken by a single speaker were tested to produce a word accuracy of 88.28%.

One of the common findings of the above studies is that demisyllable units are useful for speech recognition and are worth further research. But direct comparison of their results is not possible since they are different from one another in many ways such as target language identity, vocabulary size, quantity and quality of test data, and methods of training and test. None of them are tested over speaker independent data. Furthermore, there is no proper comparison with phone like units in terms of recognition performance as they are usually motivated by the limitations of whole word units.

5.7 Demisyllable Recogniser Construction

The main purpose of constructing demisyllable unit baseline models is of course the relative evaluation of the segmental F0 model. However, it will also be shown that regardless of the performance enhancement by exploiting F0 feature, the demisyllable model itself

[3] In fact, Japanese syllable structure is a little simpler than Korean as it has more constrained coda structure.

is better than the context-dependent phone model. As our phone model is already found to be decent in recognition performance as far as word accuracy is concerned, if the performance of demisyllable models exceeds that of phone models the demisyllable can be said to be a good recognition unit which is worth further studying.

5.7.1 Data

For appropriate cross comparisons of various models, it is important to use a common database. Accordingly, the KAIST sentence data will again be used for training and test of models. Detailed description has already been given on page 50.

5.7.2 Demisyllable generation

As the phone-level dictionary and labels have already been established following the method described in chapter 3, corresponding demisyllable labels can be created using those phone labels. To take alternative pronunciations into account, the phone level variation dictionary is used for generating a demisyllable counterpart. The core of conversion of phone strings into demisyllable strings is automatic syllabification, since syllable boundary information is indispensable. Once syllabification is complete, splitting each syllable into two demisyllables is easy.

I have already mentioned the relative simplicity of Korean syllable structure, allowing only one consonant at either onset or coda position in underlying structure. It has also been indicated that phonetic level syllable structure can be different from the underlying phonological syllable structure due to various phonological rules. This implies that phonetic syllabification is not as straightforward as phonological syllabification because the same phonotactic constraints do not apply. The syllabification process in the current experiment is conducted phonetically, since the input of the conversion is the phone level lexicon with pronunciation variants, whose items are phonetic strings after the application of phonological rules.

Phonetic syllabification can be divided into two steps: basic syllabification, which follows linguistic theories of syllabification, and resyllabification, which is a further manipulation for the convenience of practical implementation.

Basic syllabification

For the basic syllabification, I adopt traditional methods of representing syllables suggested by phonological theories such as Kahn (1976) or Clements & Keyser (1983), and their well-formedness conditions on association between items in syllable tier and segmental tier, which can be summarised as:

(5.1) a. one nucleus per syllable
b. maximal onset principle
c. syllable structure conditions

The first two items in (5.1) are widely accepted as language independent properties. The nucleus does not have to be a vowel but that is implicitly the case in Korean assuming the basic syllabification precedes any application of phonological rules including vowel deletion. The maximal onset principle is a reflection of a well-known asymmetry between onsets and codas, which is based on the observation that "there are languages lacking syllables with initial vowels and/or syllables with final consonants, but there are no languages devoid of syllables with initial consonants or of syllables with final vowels."(Jakobson (1962:526) quoted in Clements & Keyser (1983:29)). But this principle can be constrained by language specific syllabic structure conditions. It includes prohibition of a certain consonant at coda or onset position. For example, there is a constraint that /ng/ cannot be associated with onset position. As a consequence, if a segmental string is CVCVC and the second C is /ng/ as in (5.2 b) below, the syllabification will be like '[CVC] [VC]' instead of '[CV] [CVC]'. Two identical consecutive consonants in (5.2 c) are also split due to the constraint that only one segment is allowed to be associated with one onset or coda node. Through this initial syllabification, a total of 1599 syllables are generated from the KAIST data. Below are a few examples.

(5.2) Syllabification examples

a. [he] [pa] [la] [ki] 'sunflower'

b. [s a ng] [a] 'ivory'

c. [p^h u m] [m o k] 'item'

Resyllabification

The next step is refining basic syllable structure through resyllabification. The pronunciations of casual speech are different from canonical base forms because of many factors. In particular, segmental processes like deletion or weakening of sounds make it necessary to reanalyse the original syllable structure. Below are examples of phonetic hand labels for the current target database.

(5.3) **Examples of deformed syllable structure**

	original word	deformed word	glossary
(a)	p^h u m m o k	p^h m m o k	'item'
(b)	s i p s u m m i t a	s p s m t a	'hope+RESPECT+END'

Of the above cases, (a) is not difficult to resyllabify with existing syllabification criteria if the concept of syllabic consonant is introduced. That is, it can be reanalysed as two syllables like "[p^hm] [mok]" as shown in Figure 2.2 on page 20. The syllabic [m] is the nucleus of the first syllable and no criteria in (5.1) are violated.

Example 5.3 is an extreme case of deformation. Such radical vowel deletions are not very frequent but can be found in relatively fast speech, especially at the end of an intonational phrase. It is difficult to deal with this example within the framework of syllabification because of the three consecutive obstruents appearing at the beginning of the word.

To handle this problem, I adopt the Fujimura's concept of core and affix division of syllables and readjust it to suit analyses of Korean data. As most of the deformed syllables

are created through deletion or at least loss of vocality of vowel sounds, these obstruents preceding, but not immediately preceding, the vowel are separated out and regarded as affixes. Thus, the example [spsmta] can be reanalysed as:

$$[s]_{affix} + [p]_{affix} + [sm]_{syllable}\ [ta]_{syllable}$$

Contrary to English syllable affixes suggested by Fujimura, affixes of Korean can be attached to either side of core syllables. Therefore, syllables can be represented as 'affix + core syllable + affix', where occurrence of affixes is optional. Although there is no limitation on the number of affixes which can appear consecutively[4], no sequence of more than two affixes is found in the KAIST data. There are nine affixes in the current data, which are:

$$c\ k\ k^h\ p\ p^h\ t\ t^h\ s\ s'$$

It should be mentioned that the last two affixes [s] and [s'] appear most frequently, since the vowels after them are most often deleted or completely devoiced. On the other hand, the occurrence of word initial affixes of stops and affricates are relatively rare, which implies that considerable obstacles are not expected to segmental F0 extraction.

By using these affixes, the number of syllables is reduced from 1599 to 801 in the current corpus. Both the above basic syllabification and resyllabification are conducted automatically and are used for generating demisyllable label files and a lexicon.

Generating demisyllables from syllables

Once a phone string is properly syllabified according to the above method, demisyllables are easily extracted from the core syllables which can be represented by: C_iVC_f, where C_i is an optional initial consonant and C_f is an optional final consonant. Thus the representation of demisyllable derived from the corresponding syllable can be given by:

[4]For example, whispering utterances can be represented by many successive affixes.

Word and glossary	Phone string	Demisyllable string
phum-mok 'item'	p^h (u) m . m o k	p^hm m_ . mo ok
cang-nan-kam 'toy'	c a ng . n a n . k a m	ca ang . na an . ka am
a-chim 'morning'	a . c^h i m	_a a_ . c^hi im
si-cang 'market'	s (i) . c a ng	s . ca ang
thong-ci 'notice'	t^h o ng . c (i)	t^ho ong . c

Table 5.1: Examples of demisyllable strings. When a syllable begins with null onset or ends with null coda, an underline is used for a null position (eg. _a, a_). A dot is inserted between syllables. Phones surrounded by parantheses are deleted vowels.

$$C * VC* \Rightarrow (AFFIX) + (C_i)V_i + V_f(C_f) + (AFFIX)$$

where $C*$ means one or more consonants, while each of $(C_i)V_i$ and $V_f(C_f)$ stands for an initial and a final demisyllable, respectively. Table 5.1 are examples of demisyllable strings for words converted from phone strings.

5.7.3 *Demisyllable labelling*

Manually annotated label files are necessary for bootstrap training and various other purposes. As I already had phone-level hand-label files (see page 52), corresponding demisyllable labels could be automatically generated using the syllabification and syllable-to-demisyllable conversion methods described above. The procedure is as follows. First, a phone string is extracted from a label file together with beginning and end points for each phone. Second, syllabification is performed to identify the core syllable boundaries and the starting and end points of nuclei in each syllable. Finally, the period from the core syllable starting point to the midpoint of the nucleus is allocated to the initial demisyllable. Likewise, the period from the midpoint of the nucleus to the core syllable end point is allocated to the final demisyllable, which is simply the complement of the initial demisyllable. As affixes are not part of core syllables, they are irrelevant to the demisyllable segmentation and kept intact after conversion.

Figure 5.1 is an example of demisyllable segmentation converted from phone-level hand labels, aligned with its waveform signal.

In previous research, other methods have been suggested for locating the splitting point during the stable portion of the nucleus, instead of using simple nucleus midpoints. For example, Rosenberg *et al.* (1983:717) uses a point 60 msec after the start of the CV transition or the onset of voicing. Ruske (1986:49) uses a low pass filtered intensity curve, intended to simulates human loudness perception, to locate the vowel, and regards the peak of the curve as the splitting point. The 60 msec (or another fixed duration) approach, which was used for isolated word data, is not suitable for the current connected speech data in which duration of vowels is quite variable. The latter method seems to be worth trying but the peak loudness points are generally located to the left of the vowel midpoint. Thus, using this segmentation criterion will make initial demisyllables shorter than final demisyllables in general, which is not optimal to my purpose of capturing segmental F0 perturbations in initial demisyllables.

5.7.4 Lexicon

The demisyllable lexicon is constructed based on the lexicon used for phone unit recognition described on page 65. It is reasonable to assume that pronunciation variations generated for phone unit recognition will also be useful for demisyllable models. Thus, the final version of the phone lexicon is cloned and phone-to-demisyllable conversion is performed over it to make a demisyllable version of pronunciation variation dictionary. Table 5.2 shows examples of items from the demisyllable lexicon with pronunciation variation.

5.7.5 Demisyllable inventory

The above lexicon is supposed to contain all the demisyllable items to be used for training and tests. The list of demisyllables is easily generated by counting all the pronunciation

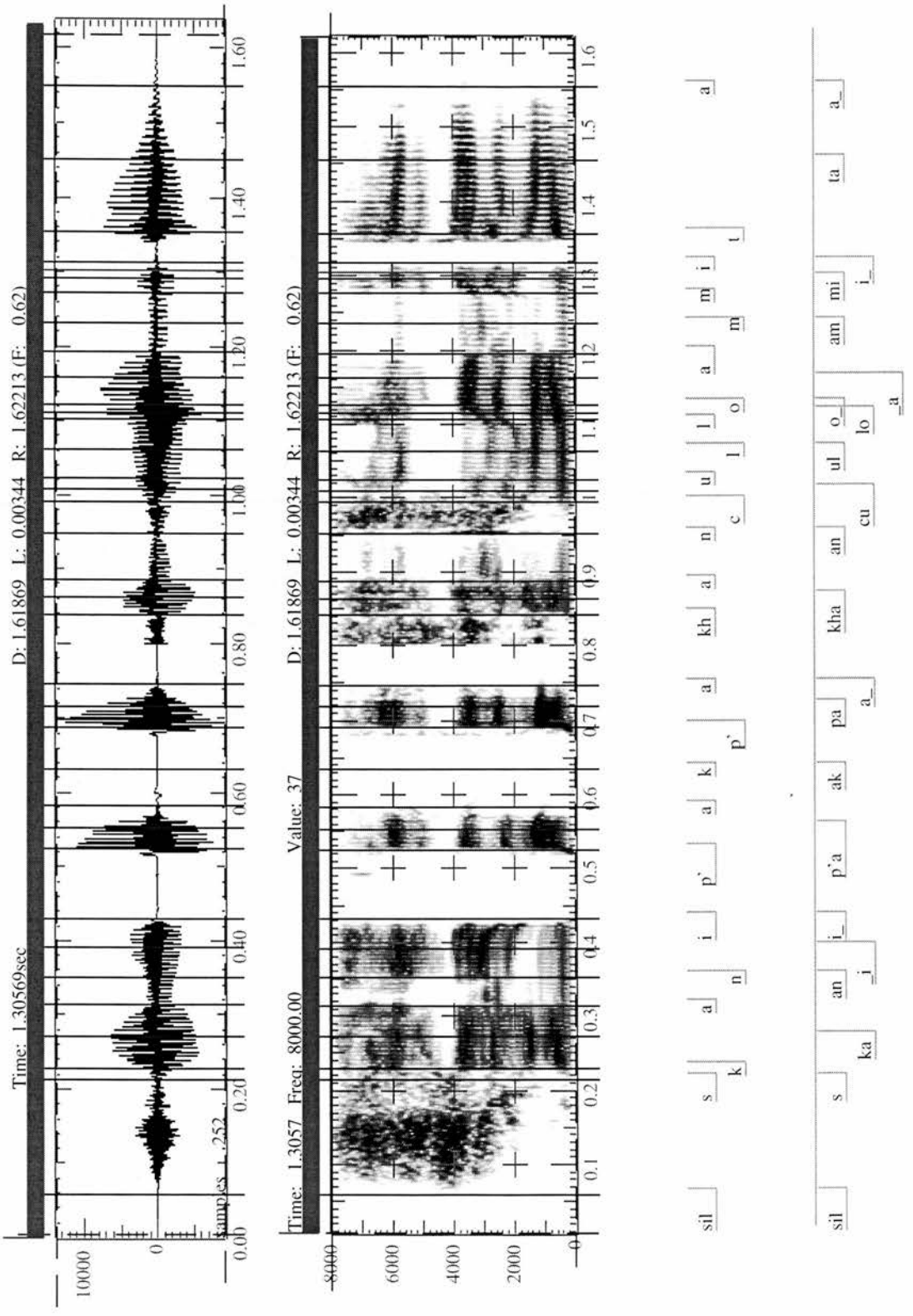


Figure 5.1: An example of demisyllable labels compared with the corresponding phone labels. The speech is *si-kan-i ppak-ppak-han cul-lo ap-ni-ta* 'We are running out of time'. The second demisyllable item, marked as [s], is an affix.

an-nae	_a a_ ne ee_
an-nae	_a an ne ee_
ceon-hwa	cv v_ na a_
ceon-hwa	cv vn ha a_
ceon-hwa	cv vn hwa wa_
ceon-hwa	cv v_ nwa wa_
iss-seup-ni-ta	_i i_ sx x_ mi i_ ta a_
iss-seup-ni-ta	_i i_ sx xm mi i_ ta a_
iss-seup-ni-ta	_i i_ sx xm ni i_ ta a_
iss-seup-ni-ta	_i i_ s' mi i_ ta a_
iss-seup-ni-ta	_i i_ s'm m_ mi i_ ta a_

Table 5.2: Examples of demisyllable lexicon items. Demisyllables without onset or coda consonants are marked as _V or V_, respectively.

Type	Structure	Number	Example
Initial demisyllable	_V	20	_a, _ya, _m
	CV	179	pa, mu, p^hm
Final demisyllable	V_	21	a_, wi_, ng_
	VC	144	ap, ing, ek
affix	C	9	s, k
silence		2	sil, sp
Total		375	

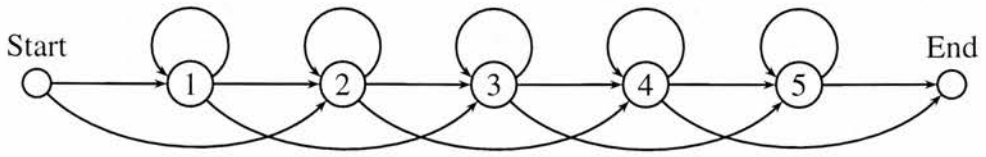
Table 5.3: Demisyllable types and numbers. Structure 'V' also stands for syllabic consonants as well as vowels.

units in the lexicon. The total number of units is 375 including 9 affixes, a silence and a short pause. The specific types of units and their numbers are shown in Table 5.3.

5.7.6 System description

Methods of feature extraction and preprocessing are basically the same as in baseline phone model construction as described on page 63.

Each demisyllable model of type CV and VC in Table 5.3 is designed to have 5 emitting states of a left-to-right HMM where transitions are permitted to skip one successive state, as in



I employed this kind of model because a demisyllable unit is usually longer than a phone unit (with 3 states). However, demisyllables composed only of a syllable nucleus part (type _V and V_ in Table 5.3), are shorter than phones as well as other demisyllables. These short demisyllables, along with silence and affixes, are represented by three state HMMs like phone units. As with the baseline phone models, a short pause model, which has just one state, is included. The single state of the pause model is tied to the centre state of the silence model and can be skipped.

Training and recognition

The demisyllable HMMs are trained with HTK. For parameter estimation, recognition, and evaluation of performance, exactly the same methods are employed as baseline phone based recognition described in chapter 3. In brief, models are initialised through using time information in hand labels, and then parameters are re-estimated by embedded training through the whole training data. A viterbi decoder is used for word recognition along with a bigram language model which is again the same one that was used in the test of the baseline phone models.

Recognition Units	Word Accuracy (%)	Sentence Accuracy (%)
Context-dependent Phones	87.93	52.15
Demisyllables	**89.58**	**54.07**

Table 5.4: Recognition performance comparison between demisyllable models and context-dependent phone models.

Results

The word accuracy obtained with the demisyllable models is 89.58%. Correct sentences are also better detected than when using phone models. Table 5.4 compares these figures for those for the context-dependent phone models. A straightforward comparison is available as the same language model and the same test tokens are used at both experiments. Considering that state-of-the-art techniques are adopted in developing the baseline phone models and their performance has already turned out to be comparable to an earlier recognition experiment based on phone like units (Choi *et al.* 1995), we can conclude that the demisyllable models are useful independent of whether enhancement using segmental F0 perturbation is achieved or not.

The result of this demisyllable recognition will be regarded as another baseline to be compared with the performance of the final models for which I attempt to exploit the segmental F0 effect.

CHAPTER 6

Connected Speech Recognition

6.1 Introduction

In this chapter, I describe the final integration of the segmental F0 effect and the demi-syllable based speech recognition system of Korean. As the experiments in the previous two chapters confirmed that (a) the segmental F0 effect is also manifest when averaging values in a relatively large amount of data, (b) segmental F0 is useful for automatic manner classification of stops and affricates, and (c) demisyllable based recognition performs better than traditional phone models at least in Korean, integration of these features is expected to enhance the overall performance of recogniser.

The most difficult problem in using segmental F0 for connected speech recognition is the interference of the utterance level F0 fluctuation. Even though we have seen that the general structure of Korean prosody is relatively concise due to lack of tonal events at the lexical level, it is obvious that segmental F0 will be affected by other prosodic factors like utterance level intonation properties such as declination of F0 downtrend or speaker pitch range variability. In the first section, I will explain how I attempt to minimise these two global influences, to increase usability of the segmental effects.

In section 6.3, more verification of the segmental F0 effect is provided. This is useful as some F0 aspects could not be checked with the isolated data of chapter 4. The main purpose is to show that segmental F0 effects are captured in connected speech data as well.

Finally, in section 6.4, a speech recognition experiment using F0 will be described. For proper comparison of performance, the previous methods will be reused wherever possible with minimal modification.

6.2 F0 Normalisation

As before, F0 extraction is based on the *get_f0* program of *ESPS* ((Entropic 1998), see page 81). For each F0 track, two types of normalisation are performed.

6.2.1 Declination normalisation

There is a global tendency for the F0 curve to decline during the course of an utterance, apart from local rises and falls. The tendency has been regarded as a language universal property of prosody as it has been widely observed cross linguistically (Pierrehumbert 1979, 1980, Ladd 1984, 1993 for English; 't Hart *et al.* 1990, Berg *et al.* 1992 for Dutch; Thorsen 1983 for Danish; Poser 1984, Kubozono 1992 for Japanese, among other studies and languages).

A similar pattern has been observed in Korean, too. Koo (1986), in an experiment similar to Pierrehumbert's (1980:118), shows how declination and prominence interact in determining the peak values. He shows that when a speaker pronounces an utterance with an emphatic stress in its second phrase, the magnitude of peak F0 in that phrase may be lower than the F0 in the first peak which is not emphatic. Other prosodic factors being assumed to have been kept constant, the declination trend accounts for this phenomenon. He also confirms that there is a consistency in the amount of declination irrespective of

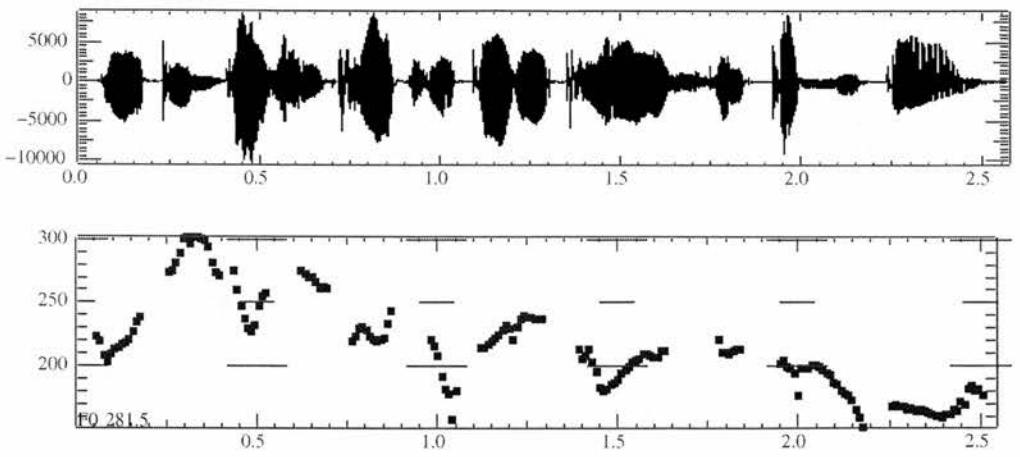


Figure 6.1: A typical example of F0 declination of a declarative sentence in the KAIST data. The sentence is *mo-teun keos-i ta ce-tae-lo toe-eo iss-eul keop-ni-ta* 'I think everything has been well done'.

the utterance duration. This means that the shorter the utterance, the steeper the declination line. Declination is present in the current data as well and Figure 6.1 is a typical example.

In order to use segmental F0 perturbation for connected speech recognition, factoring out F0 declination effect is necessary. The crucial difficulty in normalising F0 downtrend for speech recognition is that it has to be done with only the automatically extracted F0 contour of each spoken utterance, without any textual, phonological, or grammatical clues. As Ladd (1993, 1984) indicates, however, describing or normalising declination only in terms of acoustic measurements without considering more abstract phonological references is not appropriate. For example, the F0 trend during the course of a declarative sentence can be a gradual rise or a continuous level instead of the typical gradual fall, when it happens that phonologically driven pitch raising events are distributed in reverse proportion to time. In this case, though few such cases are found in current database, normalising F0 contour using a slope obtained by empirical observations may worsen the quality of F0 parameters. Therefore, a certain amount of information loss is inevitable

in the current method of normalisation and there is no claim that it fully captures the linguistic property of downtrend in intonational contours. In other words, the purpose of the normalisation is just to improve the quality of statistical parameters of segmental F0 effects.

Method

First of all, three-frame median smoothing is performed to eliminate abrupt bumps and to reduce the size of microprosodic perturbation. However, this procedure does not include smoothing or interpolation of voiceless periods as doing so has not caused any improvement in normalisation. Thus, only one or two, instead of three, frames are used for smoothing the frames close to a voiceless period.

Then, the declination slope for each utterance token is estimated by linear regression analysis in which all the F0 points, instead of only peaks and troughs of contours, are employed for quantification. Lieberman *et al.* (1985) empirically show that "all-points" regression line is a better descriptor of sentence F0 contours than either top-line or bottom-line approaches. Assuming the existence of a linear relationship between time and F0 value, the task is to find the parameters for the line. The simple linear regression model assumes that there is a line with slope β and vertical intercept α. Random deviations e are also assumed to be normally distributed and independent of one another.

$$y = \alpha + \beta x + e$$

Estimate of the parameter β for each utterance is based on all the frames with non-zero F0 values in the utterance, using the equation:

(6.1)

$$\beta = \frac{\sum x_i y_i - (\frac{\sum x_i \sum y_i}{N})}{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}$$

where x_i is index of each frame with a positive F0 value y_i, and N is the total number of frames with non-zero F0.

After the slope for a given utterance is estimated, declination normalisation is performed, effectively swinging the declination line upwards to make it horizontal, using the equation

(6.2)

$$F0(R_i) = F0(O_i) - (\beta i) \qquad 0 \leq i \leq N-1$$

where $F0(O_i)$ stands for the original F0 value of frame i and $F0(R_i)$ is the corresponding normalised value.

Figure 6.2, is an example of how the F0 downtrend has been lifted up by the Equation 6.2.

6.2.2 *Speaker normalisation*

Considering the number of speakers of both genders (35 female, 54 male speakers), the data must have a large variability of pitch range among the speakers, even among speakers of the same gender. This variation obviously affects the quality of statistical inferences. To virtually eliminate inter-speaker variability and minimise the inter-gender pitch range differences, each F0 value is readjusted on the basis of the global pitch range whose fixed values of mean and standard deviation of F0 in the male speech are calculated over all the voiced vowel tokens in the database[1]. The mean value and standard deviation of all vowels are approximately 130Hz and 25 Hz (exact values are 132 Hz

[1]Devoiced vowels are not included for calculation of pitch range.

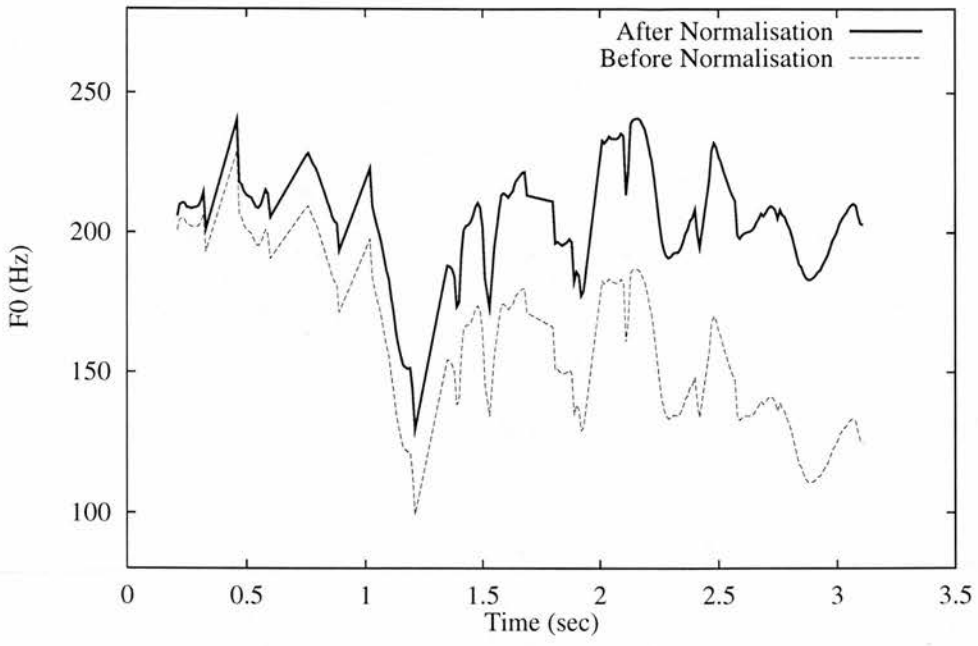


Figure 6.2: An Example of F0 Declination Normalisation.

and 24.52 Hz, respectively). Normalisation is performed to simulate this distribution. For F0 files of each speaker, values are scaled so that two standard deviations worth of values lie between 80Hz and 180Hz which indicates again two standard deviations either side of the fixed mean 130Hz[2]. In other words, 95% out of all values are made to fall between the values 80 and 180 irrespective of the individual speaker's pitch range, assuming each speaker's F0 values are normally distributed. The algorithm used for this process is shown as follows:

(6.3)

$$F0[i]_{normalised} = \left(\left(\frac{F0[i] - \mu_s}{4\sigma_s} + 0.5 \right) (High - Low) \right) + Low$$

[2]2 SD range is chosen just for instant intuitive interpretation of the normalised F0 values. Varying the range otherwise should not affect the results of the current experiment.

	Mean	StDv	Num
LAX	127.77	20.61	50025
TNS	137.90	21.40	7595
ASP	153.88	20.83	9212

Table 6.1: Overall F0 average of vowel following each obstruent class.

where μ_s and σ_s stand for the *mean* and *standard deviation* of F0 value for each utterance and Low and $High$ specify the designated pitch range. As is mentioned, the values 80 and 180 are allocated for them in current experiment.

6.3 Class Conditional Distributions

We have already seen in chapter 4 that the segmental F0 effect can be detected using automatic methods of segmentation and averaging over a fairly large amount of speech data, which contains isolated words spoken by four male speakers. As there are other factors in the current connected speech data, different from the previous isolated word data, such as a larger number of speakers and F0 declination, I have redone the statistics from chapter 4 for the new database.

Since demisyllables will be used as basic units for recognition in the next section, the extraction of F0 values will be confined to the first half of the vowel instead of the whole period. All the values below, in this section, are averaged value of the first half of each vowel period, unless specified otherwise.

6.3.1 Overall average

The overall average values in Table 6.1 confirm that the general tendency of F0 being highest after lax segments and lowest after aspirated segments is maintained in connected speech. Again the difference between "after LAX" and "after TNS" is smaller than "after TNS" and "after ASP", which makes aspirated sounds most distinguishable of three

	Mean	StDv	Num
LAX	121.68	18.71	22207
TNS	141.57	22.78	1477
ASP	160.18	20.02	3867

Table 6.2: Average F0 of post-obstruent vowels at prosodic word initial syllable.

	Mean	StDv	Num
LAX	132.62	20.76	27818
TNS	137.01	20.96	6118
ASP	149.24	20.18	5254

Table 6.3: Average F0 of post-obstruent vowels at prosodic word non-initial syllable.

types. In pairwise comparisons, the distinction between LAX and ASP is clearer than for any other combination. Though distribution of F0 after TNS falls between the two other consonantal types, resulting in a considerable overlap, its distribution is still significantly different in statistical terms ($F(2, 66738)=6412.06$, $p < 0.001$).

6.3.2 *Positional factor*

As the segmental F0 effect is most salient in prosodic word initial position (See section 2.6.1 and section 4.3.3), it is necessary to investigate whether the statistical results extracted from vowels in this position truly stand out in comparison to those at the other contexts. This is important as the speech recognition implementation will be designed to take advantage of this behaviour. Table 6.2 compared with Table 6.1 shows, as expected, that the difference between any two classes is further increased when only word initial syllables are taken into consideration.

Though the segmental effect is less salient at positions other than word initial syllables, we can see, in Table 6.3, that the effect is still maintained to a considerable degree even in non-initial positions. This means that in many, if not all, cases the effect survives even after other F0 alterating factors, indicating that the cue is generally useful all over the

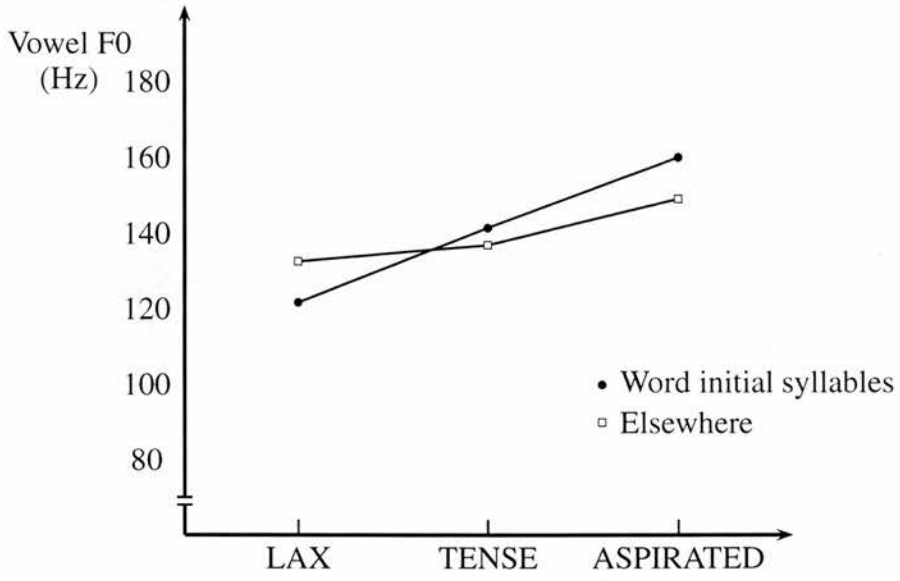


Figure 6.3: Comparison of post-obstruent F0 distribution between prosodic word initial syllable and non-initial syllable.

utterance though the degree of effect may be positionally different. However, as the main goal of the thesis is to exploit segmental F0 of word initial syllables and help identify their onset consonants, I will focus on the the occurrences of F0 perturbation at word initial position. Therefore, the results shown below through this chapter are concerned with the results for this location.

6.3.3 Hand label vs. auto label

The class conditional F0 distributions are based on phone boundary information provided by label files. As described before, those labels are automatically generated by a standard phone-unit speech recogniser and a sentence orthographic transcription for each token. Thus, it is wise to verify the reliability of auto-labelling by comparing its result with results calculated from manual labels which are assumed to be more accurate. As shown in Table 6.4, there is no considerable difference of values between hand labels (a) and auto labels (b). Statistically, none of the two F0 distributions of the same consonant manner between two different labels is significant: LAX($t = 1.28$, $p = .201$), TNS(t =

	Label Type	CLASS	Mean	St.Dv	Num
(a)	433 tokens (hand labelled)	LAX	120.31	18.65	1061
		TNS	144.04	22.25	186
		ASP	159.09	20.78	196
(b)	433 tokens (auto labelled)	LAX	121.35	19.30	1110
		TNS	144.21	23.61	146
		ASP	160.93	21.14	173
(c)	8357 tokens (auto labelled)	LAX	121.75	18.71	21146
		TNS	141.21	22.83	1291
		ASP	160.24	19.97	3671

Table 6.4: Comparison of results depending on label types. The values in row (c) are from all the training tokens without any hand labels.

	Gender	(i) LAX	(ii) TNS	(iii) ASP
(a)	54 Male Speakers	121.77	141.45	159.36
		19.13	23.17	20.41
		13423	874	2291
(b)	35 Female Speakers	121.55	141.74	161.36
		18.04	22.21	19.36
		8784	603	1576

Table 6.5: Comparison of results based on gender of speakers.

.07, p = .944), ASP(t= .84, p = .401). The results in (c) are a bit more different in values (especially after tense stops), but the difference still appears to be tolerable.

6.3.4 Speaker difference

As we have attempted to reduce speaker difference to a minimum, it is necessary to examine how this normalisation process has been reflected in the results. The gender difference appears to be properly minimised as shown in Table 6.5. No significant difference is found across genders when preceded by LAX (t = .85, p = .395), or TNS(t = .24, p = .810). The difference after ASP (t = 3.05, p = .002) also seems small and unharmful enough to conclude that the gender difference has been appropriately collapsed.

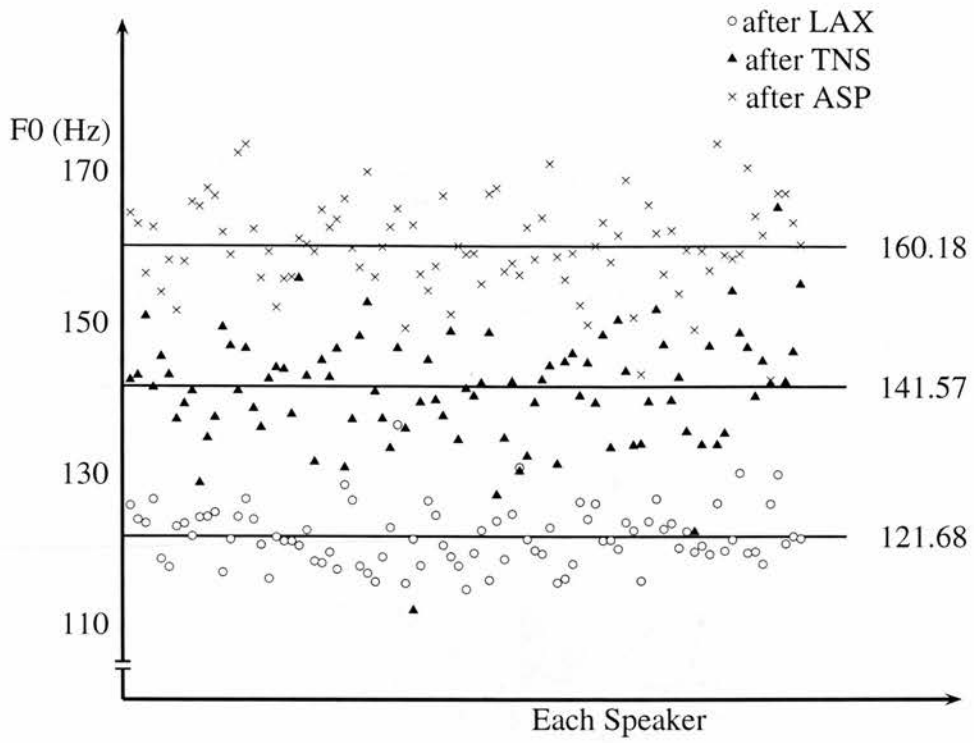


Figure 6.4: Distribution of post-consonantal F0 for each speaker. A single vertical line represents the three class-dependent average values for each speaker.

A certain degree of individual speaker variance is unavoidable even after normalisation. But it is shown in Figure 6.4 that the distinction between classes is relatively clear. Especially, no overlap is found between LAX and ASP.

6.3.5 F0 in syllabic consonants

In connected speech, vowel deletion is frequent for various reasons such as speech rate and speaker style. As discussed in section 2.5 on page 18, many cases of vowel deletion occur when followed by a sonorant consonant at coda position of the syllable. In theoretical terms, I mentioned that these cases should not cause a crucial problem in utilising the

Preceding C	Mean	StDv	Num
LAX	125.39	14.74	126
TNS	No token		
ASP	161.89	16.95	119

Table 6.6: Average F0 of post-obstruent syllabic consonants [m, n, ng, l].

segmental F0 effect since the quality of vowel, including F0, is transfered to the syllabic consonant. This section is to confirm that remark.

The F0 values of syllable final syllabic consonants [m, n, ng, l] following stops or affricates as a result of vowel deletion are averaged. Again, only the prosodic word initial position is considered. Table 6.6 is the measured values.

As expected, the F0 of syllabic consonants appears to be affected by onset consonant manner, nearly as much as the F0 of vowels are (compare with Table 6.2 on page 126). The absence of syllabic consonants after tense sounds means that vowel deletion rarely happens when preceded by tense stops or affricates, probably because of the very short aspiration period (ie., VOT) of the tense sounds compared with lax or aspirated sounds.

6.3.6 Interaction with IF0

On page 90, we have already seen that the vowel *intrinsic F0* (IF0) can be observed from a relatively large amount of data annotated automatically, in such a way that high vowels bear higher F0 than low vowels. However, the results are based on isolated word speech data. In continuous speech, the literature suggests that the effect is found to be less. Umeda (1981) even suggests that there is no consistent influence of vowel height on the F0 of vowels. But Ladd & Silverman (1984) only partially agree with her, managing to find intrinsic F0 differences, although reduced ones, in connected speech by applying more elaborate prosodic controls.

Preceding Consonant	Low V	Mid V	High V	Low-High Difference	Low-High Significance
Lax	117.04	122.78	123.55	6.51	$t = 17.08$ $p < .001$
	18.51	18.21	19.56		
	4036	9823	6788		
Tense	146.55	137.52	151.80	5.25	$t = 2.58$ $p = 0.01$
	20.78	22.68	22.52		
	416	909	149		
Aspirated	158.70	159.08	164.80	6.10	$t = 6.77$ $p < .001$
	18.96	18.96	22.32		
	1374	1569	797		
LAX-TNS Difference	29.51	14.74	28.25		
TNS-ASP Difference	12.15	21.56	13.00		

Table 6.7: Vowel intrinsic F0 effect. The three rows of each context represent, top to bottom, mean F0 (Hz), standard deviation, and number of tokens.

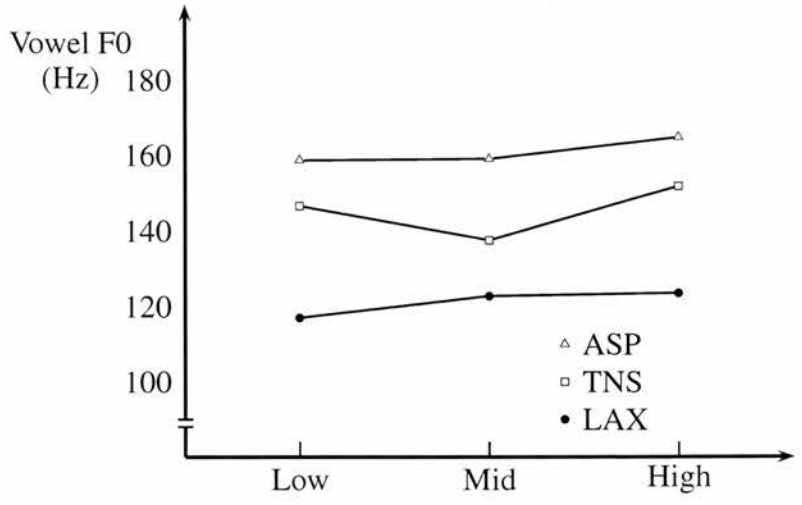


Figure 6.5: Comparison of vowel F0 depending on vowel height and consonant types.

No experimental studies, to date, have reported a Korean IF0 effect in connected speech data. Therefore, it is helpful to see if there is any significant change in the connected speech data. For a direct comparison between the IF0 effect in isolated words described in section 4.3.5 on page 89, the measurement of F0 is made by averaging the values of the whole vowel period instead of half duration.

Data type	LAX	TNS	ASP	Average
Isolated words	11.29	15.24	4.99	10.50
Connected speech	6.51	5.25	6.10	5.95

Table 6.8: The average difference of F0 between high vowels and low vowels in isolated word tokens and connected speech. Only vowels in prosodic word initial position are taken into account. The same normalisation methods were applied to both data.

The measurements presented in Table 6.7 shows that the general tendency that high vowels bear higher F0 than lax vowels is still observed in the connected speech and it is statistically significant. Note, however, that the difference caused by vowel height is relatively small in comparison to the difference caused by the type of the preceding consonant. F0 difference between low vowels and high vowels is 6.51, 5.25, 6.10 Hz for the classes Lax, Tense, and Aspirated, respectively. Table 6.8 is given to compare these figures with those of isolated words extracted from the Table 4.8 on page 90.

The comparison confirms that in general the IF0 effect is considerably weaker in connected speech agreeing with the results in Ladd & Silverman (1984). The exceptional case is when the preceding consonant is an aspirated sound. Here the IF0 effect is slightly increased although that is the case in which it was least prominent for the isolated words. Note also that, in either type of data, IF0 difference after aspirated sounds is smaller than after lax sounds. This result appears not in accordance with the previous observations that IF0 difference is more conspicuous when the vowel is in the pitch accented syllable (Terken 1995:111). Consequently, it can be inferred that segmental F0 does not affect vowel intrinsic F0 considerably, though further investigation seems necessary.

As it is the case that generally the IF0 effect has decreased in connected speech, the influence of IF0 on the consonantal effect is also expected to be less than in isolated words. In Table 6.7, there is still a considerable distance between the F0 value of high vowels preceded by lax obstruents (123.55), and the F0 of low vowels preceded by aspirated

Context	After [s]	After [s']
P-word Initial Syllable	157.61	149.22
	19.99	17.65
	4927	144
P-word Final Syllable	143.16	138.65
	20.80	22.04
	8734	2228

Table 6.9: Segmental F0 after alveolar fricatives.

obstruents (158.70). This means that the vowel height factor, though preserved in Korean connected speech as well, does not appear to preempt the consonantal effect.

6.3.7 F0 after alveolar fricatives

In section 2.4, I mentioned that there was a controversy on whether a Korean alveolar fricative /s/ should be classified as lax or aspirated, while /s'/ has been unanimously classified as tense. Segmental F0 gives a useful clue for resolving this controversy, as it is particularly good at the lax/aspirated distinction. As shown in Table 6.9, the vowel F0 after [s] is higher than after [s'] irrespective of the syllable position. The difference between two classes is also highly significant ($F(1, 5069)=24.77$, $p < 0.001$, for prosodic word initial position, and $F(1, 10960)=81.20$, $p < 0.001$, for elsewhere). It is obvious that the F0 distribution after [s] is similar to that after aspirated stops and affricates rather than after lax stops in Table 6.2 and 6.3 on page 126. Thus, I support the argument that Korean alveolar fricatives are classified into aspirated and tense categories, instead of lax and tense categories.

It is not likely that using the segmental F0 for fricative distinction will considerably increase the recognition performance. The frequency of [s'] is very small at the prosodic word initial position where the segmental effect is most helpful. Furthermore, we saw in

section 2.7 on page 32 that the recognition accuracy of fricatives is already quite higher compared to stops and affricates.

However, the result of this investigation is important since a generalisation of the segmental F0 effect to nearly all Korean obstruents can be achieved.

6.4 Implementation

6.4.1 Prosodic word as a unit of AP constituency

On page 85, we saw that the segmental F0 effect is mainly useful at the AP initial position. However, it is very difficult to use this positional information explicitly in a speech recognition task. The crucial difficulty is that AP is not easy to detect automatically. Recently, Lee & Song (2000) reports that 73% of Korean APs are detected using a *dynamic time warping* technique (Owens 1993:140). But their experiment was only on a small amount of read speech data obtained from a single speaker. The task will be a lot more difficult in a speaker independent ASR system. In fact, the crucial problem is that automatic AP detection has to be done using the F0 contour, as AP is mainly defined by intonational structure (See page 23). Thus it is quite probable that the segmental F0 effect will eventually hamper the detection of APs, especially when aspirated sounds are involved in the AP. Thus, using segmental F0 effect in speech recognition after automatic AP demarcation is not plausible.

Thus we need to adopt a simple operational definition of AP constituency. We have noted that an AP is in principle composed of one or more prosodic words. However, I found that a vast majority of APs in the current data consist of only one prosodic word, which suggests that it should be possible to identify AP initial position with prosodic word initial position, and just distinguish word initial and word non-initial demisyllables.

Feature Type	Feature	Dimension
MFCC	Static	12
	Delta	12
	Acceleration	12
Energy	Static	1
	Delta	1
	Acceleration	1
F0	Static	1
	Delta	1
Total	-	41

Table 6.10: Features.

6.4.2 *Demisyllable inventory*

The demisyllable inventory for the present experiment is initially derived from the baseline inventory of chapter 5. However, when the type of word initial demisyllable is CV, and the consonant is a stop or affricate, it is treated as a different demisyllable from the corresponding non-word-initial demisyllable. For example, a demisyllable *pa* can be split up into two separate units: word initial, and word non-initial. Theoretically, the number of new units may increase by 216 (12 stops and affricates x 18 vowels). But only 99 more demisyllables are added to the inventory as there are many CV combinations which do not occur within the corpus. In particular, the number of demisyllables with 'C + diphthong' combination was relatively small.

Consequently, the total number of demisyllable units in the current experiment is found to be 473 including 9 affixes, a silence, and a short pause.

6.4.3 *Feature*

As with the demisyllable baseline model construction, 12 MFCC, energy and their first and second derivatives are extracted as parameters. F0 values are added to those existing

features. Again, an ESPS program (get_f0) is used to extract F0 values. Delta values for F0 are also calculated. These features are summarised in Table 6.10.

6.4.4 *Training and recognition*

Basically, the same methods of training and recognition are employed as in the demisyllable baseline model. Bootstrapping with hand labelled information, Baum-Welch re-estimation, and embedded retraining with whole train data are conducted in turn.

As the number of units has increased considerably due to separation of prosodic word initial demisyllables, a relative sparsity of training data can be a problem. To cope with this, a parameter tying technique, so-called *data-driven clustering*, is adopted (Young *et al.* 1996:151). Firstly, similar units are classified as the domain of parameter clustering on the basis of linguistic intuition. This is to prevent generation of clusters with very little associated training data. Then, acoustic parameters of each unit within the same class are compared. The acoustic distance between the same level states of two units is calculated and if it is smaller than an artificially set threshold, the two states are merged and made to share the same parameters. For example, demisyllables such as 'pwa' and 'pa' can be judged to be acoustically very similar and the former has a very small number of training tokens. Then they are classified in the same group and the parameters of each state of the two units are compared to judge whether they are similar enough (ie., within threshold) to be merged. The threshold is adjusted heuristically by monitoring recognition performance and reduction size of the number of states. An HTK program HHEd and its internal command TC are used for this process.

All the other methods of training and recognition are equivalent to those used in the demisyllable baseline construction (page 116).

Model	Word Accuracy
triphone	87.93
demisyllable	89.58
demisyllable + F0	91.03

Table 6.11: Comparison of performance.

6.4.5 *Evaluation*

The word accuracy achieved in this experiment is 91.03%. The comparison of word accuracy with other models, in Table 6.11 shows that the segmental F0 models are slightly better than plain demisyllable models as well as context-dependent phone models. Though the increase of word accuracy is small (1.45% from the demisyllable baseline, 3.10% from phone baseline) it represents a 14% reduction in error rate, and it is believed to be meaningful since only minimal changes were made to the previous system and factors other than segmental F0 in the experiment were kept constant. The number of test data tokens is also relatively large (2073 utterances), so that such improvement is unlikely to have been achieved by chance.

Significance of improvement

As the extent of word accuracy increase is small (1.5%) a statistical significance test will be helpful to confirm that it is not achieved by chance. To generate values for this purpose, recognition accuracy for each utterance token is calculated separately. As the number of test utterances is 2073, a set composed of this many accuracy values is obtained for both the demisyllable baseline model and the demisyllable F0 model. Thus, a pair of data sets is obtained for significance testing. However, the values in each set are not normally distributed because many of the values are simply '100%' whether for the baseline model or the F0 model [3]. This means that the most frequently used tests such as the t-test

[3]The *skewness* and *kurtosis* for each distribution were -3.47, 20.48 for baseline model, and -3.57, 21.50 for F0 model.

or F-test are not suitable here since one of their basic assumptions is the normality of data distribution. Instead, I performed a simple statistical inference test called the 'sign test' (Clarke & Cooke 1992:175, 237) in which the number of positive and negative signs obtained by pairwise comparison of sample items are used for testing significance. The result of the test shows that the difference between the two data sets is statistically significant at the 95% confidence level ($p = 0.0401$).

A supplementary test

When we closely examine the demisyllable models with and without F0 implementation, a suspicion may arise in a way that the improvement of performance could have been caused by an artifact of an experimental design. That is to say, I have treated word initial stops and affricates separated from those located in any other position, to take a maximum advantage of the segmental F0 effect which is more prominent in word initial position. An hypothesis is that this splitting of demisyllable inventory, instead of segmental F0, may be responsible for the improvement. To check this out, I conducted another recognition test in which F0 parameters were not used at all but still the consonants in word initial position are treated separately. All other experimental design was same as before.

The word accuracy from this test was 89.33%. It turns out that no improvement of recognition performance has been made in comparison to the accuracy of baseline model, which was 89.58%. Therefore, it can be inferred that the improvement shown in Table 6.11 was not caused by the separation of obstruent categories into 'word initial' and 'elsewhere'.

In conclusion, it is legitimate to state that implementation of segmental F0 helps to increase word accuracy in Korean ASR. Considering that the size of total error is approximately 10%, and that there are, as we have seen in section 2.7 on page 32, various other sources causing errors like consonantal place confusion which are not dealt with in this study, the improvement achieved by the current method can be judged to be meaningful.

CHAPTER 7

Conclusion

In this final chapter, I will give a summary of major findings in this research. Then I will indicate which components of the recognition process with segmental F0 model need to be further studied in order to produce better results.

7.1 Summary of Findings

This thesis has shown that careful phonetic analysis, of a fairly traditional sort, can contribute to statistically based speech recognition by guiding the choice of units whose behaviour is to be statistically modelled.

Various language particular characteristics of spoken Korean make the segmental F0 effect salient in this language. The phonotactic constraint which prohibits consonant clusters reinforces consonant-vowel coherence making it easy for a consonant to transfer its articulatory characteristic to the following vowel causing manner specific F0 fluctuation. Furthermore, there is no distinctive lexical stress or tone to mask segmental F0 effects. Other utterance level F0 influencing factors such as phrasal tones are also located in positions where conflict with the segmental F0 effects is minimised.

Segmental F0 has proved to be useful in speech recognition of Korean. The high F0 after aspirated and tense obstruents and low F0 after lax obstruents do contribute to the identification of consonant manners, especially at the prosodic word initial position. Though the increase of accuracy from the baseline model is fairly small, considering that there are various other sources of error and that the improvement is verified by significance testing, the implementation of segmental F0 in speech recognition has been successful.

Automatic statistical analyses on a relatively large amount of data can appropriately model the segmental F0 effect. Many properties related to segmental F0, which have been discovered and verified by carefully controlled phonetic experiments and are accepted widely by researchers to be true properties, are also captured by statistical approximation. Instead of minute controls of context in data which require high-level linguistic information, automatic controls on F0 which can be performed on surface level acoustic signals were provided to reduce the F0 variability in data. The manifestation of F0 after such automatic processing and normalisation has been a firm basis for further application of F0 in speech recognition.

Demisyllables were used as basic recognition units. The immediate motivation for employing the new units was convenience in combining F0 parameters with standard spectral features. Even apart from this advantage, however, I found that a speech recogniser with demisyllable units works better than a baseline recogniser with context dependent phone units. To take full advantage of segmental F0 effects, word initial demisyllables were distinguished from non-word initial ones, since F0 effects more pronounced word initially, while spectral parameters play a more important role in distinguishing stop and affricate consonants non-initially.

Apart from the F0 implementation, I enhanced speech recognition performance by modifying the lexicon to contain various alternative pronunciations. A large number of variants of each canonical pronunciation were initially generated using standard phonological rules and new ones based on the performance of the baseline recogniser on training data. Then overgeneration was constrained by a data-driven selection process, by which

less frequent variants, from the point of view of the speech recogniser, were systematically discarded through a heuristic adjustment of probability thresholds. A considerable improvement in word accuracy was achieved with this method.

7.2 Further Work

Different data

The database used in my experiments is fairly well organised in the sense that the vocabulary size is moderate (2920), a fairly large number of speakers in both genders are involved (89), and the utterances are relatively natural. However, one cannot take it for granted that the same enhancement can be achieved with other data. Therefore, replication of these experiments with different databases would provide useful confirmation. In particular, data consisting of spontaneous utterances would be valuable, as some F0 patterns are known to be different between read and spontaneous speech (Kowtko 1997). This, of course, requires the construction of a relevant Korean database.

Normalising F0 Declination

The weakest point of F0 implementation for continuous speech recognition in the thesis is normalisation for declination, as it is quite likely that a fair amount of information has been lost during this process. The main part of the current method is finding a best-fit line to F0 values in voiced frames across the whole utterance. But for utterance tokens with more than one intonational phrase, this method is faulty because a single slope cannot characterise the declination of all prosodic units. Thus, in order to localise the domain of normalisation to a more sensible portion, breaking an utterance into subparts first would be necessary. Other prosodic cues like global F0 or pauses, as previous studies suggest (see Introduction), will be helpful for this.

Another important point is that ultimately declination at a given point in the utterance needs to be estimated from values during the past time span instead of those of the whole

utterance. This is essential for a recogniser to perform real time recognition without having to wait until the end point of the utterance before starting to run the HMMs.

In a speech recognition experiment with a database with only isolated words, for which declination normalisation is unnecessary, I achieved a better word accuracy increase (2.37%), from a baseline result with only spectral features (94.35%) to the result with the F0 assisted model (96.72%). This performance improvement compared with what I achieved from continuous speech recognition (1.45%) suggests that more improvement might be possible with better normalisation of declination. Though there will be a restriction on improving normalisation due to its association with abstract level linguistic information which is hard to capture from the acoustic signal alone, comprehensive studies of Korean intonation and accents will help find a better automatic algorithm for reducing errors.

F0 processing

A better F0 extraction algorithm will improve the reliability of segmental F0 modelling. Silverman (1987:2.17) found that his automatic F0 algorithm was least accurate around the boundary between syllable onset consonant and starting point of the vowel nucleus. As the segmental F0 effect persists for some time into the vowel, averaging at least half the duration of each vowel in the measurement of F0 value is expected to have decreased the effect of such automatic detection errors around the beginning of the vowels in my experiments. However, making fewer errors in the first place would contribute to improved statistical modelling by allowing us to concentrate on the very beginning of the vowel where the segmental effect is most prominent. A robust F0 detection algorithm is also necessary for extracting F0 from data created in adverse conditions.

Another major difficulty in using F0 information is interaction of factors. Separating out local F0 perturbation from other factors is not easy even for such a target language as Korean which is known to have a particularly prominent segmental F0 effect. Means

need to be found to minimise the loss of information caused by this factor interference. This will lead to successful application of segmental F0 to many other languages.

Implementation

Combining segmental F0 with other methods of F0 exploitation is also desirable. As mentioned in the Introduction, there are other uses of F0: disambiguation of utterance meaning, demarcating prosodic boundaries, and improving language model using utterance type detection. As most of these studies report positive results and as most of them depend on F0 as the crucial information source, better performance with relatively less effort for information extraction is expected.

7.3 Final Remarks

Pitch in human spoken language plays a very important role in communication. Thus, without proper analysis of its acoustic correlate, F0, automatic speech recognition cannot adequately simulate human perception. Recently, there has been an increase in the number of studies on F0, as well as other prosodic cues, for speech recognition. However, there have been few reports on the use of F0 for Korean speech recognition, even fewer on the use of segmental F0. Thus, I hope the work presented can be a starting point and benchmark for other related work in automatic speech recognition of Korean.

Bibliography

ABRAMSON, ARTHUR S., & LEIGH LISKER. 1971. Voice timing in Korean stops. In *Proceedings of teh Seventh International Congress of Phonetics*, Montreal, Canada.

BECKMAN, MARY E., & JANET B. PIERREHUMBERT. 1986. Intonational structure in Japanese and English. *Phonology Yearbook 3* 255–309.

BELLMAN, R. E. 1957. *Dynamic Programming*. Princeton, New Jersey, USA: Princeton University Press.

BERG, R. VAN DEN, C. GUSSENHOVEN, & T. RIETVELD. 1992. Downstep in Dutch: implications for a model. *Papers in Laboratory Phonology* 2.335–359.

BEULEN, K., S. ORTMANNS, A. EIDEN, S. MARTIN, L. WELLING, J OVERMANN, & H. NEY. 1998. Pronunciation modelling in the RWTH large vocabulary speech recognizer. In (Strik *et al.* 1998), 13–16.

CHOI, I. J., O. W. KWON, J. R. PARK, Y. K. PARK, D. Y. KIM, H. Y. JEONG, & C. K. UN. 1995. On the development of a large-vocabulary continuous speech recognition system for the Korean language. *The Jounal of the Acoustical Society of Korea* 14(5).44–50. (in Korean).

CHOMSKY, NOAM, & MORRIS HALLE. 1968. *The Sound Pattern of English*. Harper & Row.

CLARKE, G. M., & D. COOKE. 1992. *A basic course in statistics*. Edward Arnold, 3rd edition.

CLARKSON, PHILIP, & RONALD ROSENFELD. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *European Conference on Speech Communication and Technology 97*, volume 5, 2707–2710.

CLEMENTS, GEROGE N., & SAMUEL JAY KEYSER. 1983. *CV Phonology: A Generative Theory of the Syllable*. Linguistic Inquiry Monograph. The MIT Press.

DE PIJPER, J. R. 1983. *Modelling British English intonation : an analysis by resynthesis of British English intonation*. Dortrecht, Foris.

DEVOER, JAY, & NICHOLAS FARNUM. 1999. *Applied Statistics for Engineers and Scientists*. Duxbury.

DUMOUCHEL, P. 1994. Suprasegmental features and continuous speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing 94*, volume 2, p. 177.

DUMOUCHEL, P., & D. O'SHAUGHNESSY. 1993. Prosody and continuous speech recognition. In *European Conference on Speech Communication and Technology 93*, volume 3, 2195–2198, Berlin.

ENTROPIC. 1998. *ESPS/Waves+ with EnSig 5.3*. Version 5.3.

EUN, J. K., KIM, & PARK. 1989. Research on the development of speech recognition systems for korean language. Technical Report 1, Korea Advanced Institute of Science and Technology, Seoul, Korea. (in Korean).

FEGYÓ, TIBOR, & PÉTER TATAI. 1999. Multi-lingual speech recognition based on demi-syllable subword units. In *European Conference on Speech Communication and Technology 99*, volume 2, 867–870.

FERGUSON, J. D. 1980. Variable duration models for speech. In *Proceedings of teh Symposium on the Application of Hidden Markov Models to Text and Speech*, ed. by J. D. Ferguson, 143–179, Princeton, New Jersey.

FERREIROS, J., J. MACÍAS-GUARASA, J. M. PARDO, & L. VILLARRUBIA. 1998. Introducing multiple pronunciations in Spanish speech recognition systems. In (Strik *et al.* 1998), 29–33.

FORNEY, G. D. 1973. The Viterbi algorithm. *Proc. IEEE* 61.268–278.

FUJIMURA, OSAMU. 1976. Syllables as concatenated demisyllables and affixes. *The Jounal of the Acoustical Society of America* 59(Suppl.1).S55. (abstract).

FUJIMURA, OSAMU, & DONNA ERICKSON. 1997. Acoustic phonetics. In *The handbook of phonetic sciences*, ed. by William J. Hardcastle & John Laver, chapter 3, 65–115. Blackwell.

GANAPATHIRAJU, A., V. GOEL, J. PICONE, A. CORRADA, G. DODDINGTON, K. KIRCHOFF, M. ORDOWSKI, & B. WHEATLEY. 1997. Syllable - a promising recognition unit for lvcsr. In *roceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 207–214, Santa Barbara, California, USA.

GANDOUR, JACK. 1974. Consonant types and tone in siamese. *Journal of Phonetics* 2.337–350.

GOLDSMITH, J., 1976. *Autosegmental Phonology*. MIT dissertation.

HAGGARD, MARK, STEPHEN AMBLER, & MO CALLOW. 1970. Pitch as a voicing cue. *The Jounal of the Acoustical Society of America* 47.613–617.

HALLE, M., & K. N. STEVENS. 1971. A note on laryngeal features. In *Quarterly Progress Report of the Research Laboratory of Electronics*, volume 101, 198–213. MIT.

HAN, J. I., 1996. *The Phonetics and Phonology of Tense and Plain Consonants in Korean*. Cornell University dissertation.

HAN, MIEKO S., & R. S. WEITZMAN. 1967. Studies in the phonology of asian languages v: Acoustic features in the manner-differentiation of Korean stop consonants. Technical report, University of Southern California.

HAN, MIEKO S., & R. S. WEITZMAN. 1970. Acoustic features of Korean /P, T, K/, /p, t, k/ and /p^h, t^h, k^h/. *Phonetica* 22.112–128.

HARDCASTLE, W. J. 1973. Some observations on the tense-lax distinction in initial stops in Korean. *Journal of Phonetics* 1.263–272.

HIROSE, H., C. Y. LEE, & T. USHIJIMA. 1974. Laryngeal control in Korean stop production. *Journal of Phonetics* 2.145–152.

HIROSE, KEIKICHI. 1996. Disambiguating recognition results by prosodic features. In (Sagisaka *et al.* 1996), chapter 20, 327–342.

HIRST, DANIEL. 1983. Structures and categories in prosodic representations. In *Prosody: Models and Measurements*, ed. by A. Cutler & D. R. Ladd, chapter 8, 93–109. Berlin: Springer Verlag.

HOMBERT, JEAN-MARIE. 1978. Consonant types, vowel quality, and tone. In *Tone: A linguistic survey*, ed. by Victoria Fromkin, 77–111. Academic Press.

HOUSE, ARTHUR S., & GRANT FAIRBANKS. 1953. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Jounal of the Acoustical Society of America* 25(1).105–113.

HU, ZHIHONG, JOHAN SCHALKWYK, ETIENNE BARNARD, & RONALD A. COLE. 1996. Speech recognition using syllable-like units. In *International Conference on Spoken Language Processing 96*, volume 2.

HUH, W. 1991. *Korean Phonology*. Seoul, Korea: Sam Munhwasa. (in Korean).

HUNT, ANDREW J. 1996. Trining prosody-syntax recognition models without prosodic labels. In (Sagisaka *et al.* 1996), chapter 20, 309–325.

IVERSON, GREGORY K. 1983. Korean s. *Journal of Phonetics* 11.191–200.

J. PAE, J. SHIN, & D. KO. 1999. Some acoustical aspects of Korean stops in various utterance positions :focusing on their temporal characteristics. *Korean Journal of Speech Sciences* 5(2).139–159. (in Korean).

JAKOBSON, R. 1962. *Selected Writings 1: Phonological Studies*. The Hague: Mouton, 2nd expansion edition.

JANG, TAE-YEOUB. 2000. Fundamental frequency in manner differentiation of Korean stops and affricates. *Speech Sciences* 7(1).217–232. The Korean Association of Speech Sciences.

JANG, TAE-YEOUB, MINSUCK SONG, & KIYEONG LEE. 1998. Disambiguation of Korean utterances using automatic intonation recognition. In *International Conference on Spoken Language Processing 98*, volume 3, 603–606.

JELINEK, F. 1990. Self-organized language modeling for speech recognition. In *Readings in Speech Recognition*, ed. by Alex Waibel & Kai-Fu Lee, 450–506. Morgan Kaufmann.

JUN, SUN-AH, 1993. *The phonetics and phonology of Korean prosody*. The Ohio State University dissertation.

JUN, SUN-AH. 1996. Influence of micorprosody on macroprosody: a case of phrase initial strengthening. *UCLA Working Papers in Phonetics* 92.97–116.

JUN, SUN-AH. 1998. The accentual phrase in the Korean prosodic hierarchy. *Phonology* 15(2).189–226.

KAGAYA, RYOHEI. 1971. Laryngeal gestures in Korean stop consonants. Technical Report 5, Research Institute of Logopedics and Phoniatrics, University of Tokyo.

KAGAYA, RYOHEI. 1974. A fibersocopic and acoustic study of the Korean stops, affricates and fricatives. *Jounal of Phonetics* 2.161–180.

KAHN, DANIEL, 1976. *Syllable-based generalizations in English Phonology*. MIT dissertation.

KIECZA, DANIEL, TANJA SCHULTZ, & ALEX WAIBEL. 1999. Data-driven determination of appropriate dictionary units for korean lvcsr. In *International Conference on Speech Processing 99*, volume 1, 323–327.

KIM, CHIN-WU. 1965. On the autonomy of the tensity feature in stop classification (with special reference to Korean stops). *Word* 21(3).339–359.

KIM, CHIN-WU. 1970. A theory of aspiration. *Phonetica* 21.107–116.

KIM, KYUNGNYUN, 1968. F0 variations according to consonantal environments. Ms. University of California at Berkeley.

KING, SIMON A., 1998. *Using Information above the Word Level*. University of Edinburgh dissertation.

KOHLER, KLAUS J. 1982. F0 in the production of lenis and fortis plosives. *Phonetica* 39.199–218.

KOHLER, KLAUS J. 1985. F0 in the perception of lenis and fortis plosives. *The Jounal of the Acoustical Society of America* 78(1).21–32.

KOHLER, KLAUS J. 1986. Preplosive f0 in the perception of /d/-/t/ in English. In *Montreal Symposium on Speech Recognition*, 34–35. McGill University.

KOMPE, RALF. 1997. *Prosody in Speech Understanding Systems*. Lecture Notes in Artificial Intelligence. Berlin: Springer Verlag.

KOO, HEE SAN, 1986. *An Experimental Acoustic Study of the Phonetics of Intonation in Standard Korean*. University of Texas at Austin dissertation.

KOREMAN, JACQUES, WILLIAM J. BARRY, & BISTRA ANDREEVA. 1997. Relational phonetic features for consonant identification in a hybrid ASR system. *Phonus* 3.83–109.

KOWTKO, JACQUELINE. 1997. The function of intonation in spontaneous and read dialogue. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 286–289.

KUBOZONO, HARUO. 1992. Modeling syntactic effects on downstep in Japanese. *Papers in Laboratory Phonology* 2.368–397.

LADD, D. ROBERT. 1984. Declination: a review and some hypotheses. *Phonology Yearbook* 1.53–74.

LADD, D. ROBERT. 1993. On the theoretical status of "The baseline in modelling intonation". *The Jounal of the Acoustical Society of America* 36(4).435–451.

LADD, D. ROBERT. 1996. *Intonational Phonology*. Cambridge University Press.

LADD, D. ROBERT, & KIM E. A. SILVERMAN. 1984. Vowel intrinsic pitch in connected speech. *Phonetica* 41.31–40.

LADEFOGED, PETER. 1973. The features of the larynx. *Journal of Phonetics* 1(1).73–83.

LAVER, JOHN. 1994. *Principles of Phonetics*. Cambridge University Press.

LEA, W. A. 1980. Prosodic aids to speech recognition. In *Trends in Speech Recognition*, ed. by W. A. Lea, 166–205. Prentice Hall.

LEA, WAYNE A. 1977. Acoustic correlates of stress and juncture. *Studies in Streess and Accent* 4.83–119.

LEE, C. Y., & T. S. SMITH. 1972. A study of subglotal air pressure in Korean stop consonants. *The Jounal of the Acoustical Society of America* 51(1). (abstract).

LEE, H. 1996. *Korean Phonetics*. Seoul, Korea: Taehaksa. (in Korean).

LEE, H. B. 1973. A phonetic study of the accent in Korean. *Mullidaehakpo (Seoul National University)* 19.1–16. (in Korean).

LEE, HO-YOUNG, 1990. *The Structure of Korean Prosody*. University College London dissertation.

LEE, HYUCK-JOON. 1998. Non-adjacent segmental effects in tonal realization of accentual phrase in Seoul Korean. In *International Conference on Spoken Language Processing 98*, volume 3, 623–626.

LEE, KIYOUNG, & MINSUCK SONG. 2000. Automatic detection of intonational and accentual phrases in Korean standard continuous speech. *Speech Sciences* 7(2).209–224. (In Korean).

LEHISTE, ILSE, & GORDON E. PETERSON. 1961. Some basic considerations in the analysis of intonation. *The Jounal of the Acoustical Society of America* 33(4).419–425.

LEVINSON, S. E. 1986. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech and Language* 1.29–45.

LIBERMAN, ALVIN M. 1996. Perceptual equivalence of two acoustic cues for stop-consonant manner. In *Speech: A Special Code*, ed. by Alvin M. Liberman, chapter 20, 371–383. The MIT Press.

LIEBERMAN, P., W. KATZ, A. JONGMAN, R. ZIMMERMAN, & M. MILLER. 1985. Measures of the sentence intonation of read and spontaneous speech in American English. *The Jounal of the Acoustical Society of America* 77(2).649–657.

LISKER, L., & A. ABRAMSON. 1964. A cross-language study of voicing in initial stops: Acoustic measurements. *Word* 20.384–422.

LÖFQVIST, A. 1975. Intrinsic and extrinsic fo variations in swedish. *Phonetica* 31.228–247.

MOHR, B. 1971. Intrinsic variations in the speech signal. *Phonetica* 23.65–93.

NAKAI, MITSURU, HARALD SINGER, YOSHINORI SAGISAKA, & HIROSHI SHIMODAIRA. 1996. Accent phrase segmentation by f0 clustering using superpositional modelling. In (Sagisaka *et al.* 1996), chapter 20, 343–359.

NESPOR, M., & I. VOGEL. 1986. *Prosodic Phonology*. Dordrecht: Foris Publication.

NEY, H., U. ESSEN, & R. KNESER. 1994. On structuring probablistic dependencies in stochastic language modelling. *Computer Speech and Language* 8(1).1–38.

OH, MIRA. 1999. Korean prosodic structure and focus. In *International Conference on Phonetic Science 99*, 1517–1520, San Francisco.

OHALA, JOHN J. 1978. Production of tone. In *Tone: A linguistic survey*, ed. by Victoria Fromkin, 5–39. Academic Press.

OHDE, RALPH N. 1984. Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Jounal of the Acoustical Society of America* 75(1).224–320.

OWENS, F. J. 1993. *Signal Processing of Speech*. The MacMillan Press.

PARK, HANSANG. 1999. The phonetic nature of the phonological contrast between the lenis and fortis fricatives in Korean. In *International Conference on Phonetic Science 99*, 425–428.

PARK, JONG RYEAL, OH WOOK KWON, DO YEONG KIM, IN JEONG CHOI, HO YOUNG JEONG, & CHONG KWAN UN. 1995. Speech data collection for Korean speech recognition. *The Journal of the Acoustic Society of Korea* 14(4).74–81. (in Korean).

PETERSEN, REINHOLT. 1986. Perceptual compensation for segmentally conditioned fundamental frequency perturbation. *Phonetica* 43.31–42.

PETERSON, G. E., & H. L. BARNEY. 1952. Control methods used in a study of the vowels. *The Jounal of the Acoustical Society of America* 24.175–184.

PIERREHUMBERT, JANET. 1979. The perception of fundamental frequency declination. *The Jounal of the Acoustical Society of America* 66(2).363–369.

PIERREHUMBERT, JANET, 1980. *The phonology and phonetics of English intonation*. MIT dissertation.

POSER, WILLIAM J., 1984. *The phonetics and phonology of tone and intonation in Japanese*. MIT dissertation.

PRICE, P. J., M. OSTENDORF, S. SHATTUCK-HUFNAGEL, & C. FONG. 1991. The use of prosody in syntactic disambiguation. *The Jounal of the Acoustical Society of America* 90(6).2956–2970.

RABINER, L. R., & B. H. JUANG. 1993. *Fundamentals of Speech Recognition*. New Jersey: Prentice Hall.

RILEY, M., W. BYRNE, M. FINKE, S. KHUDANPUR, A. LJOLJE, J. MCDONOUGH, H. NOCK, M. SARACLAR, C. WOOTERS, & G. ZAVALIAGKOS. 1998. Stochastic pronunciation modelling from hand-labelled phonetic corpora. In (Strik *et al.* 1998), 109–116.

ROSENBERG, A. E., L. R. RABINER, J. G. WILPON, & D. KAHN. 1983. Demisyllable-based isolated word recognition. *IEEE transactions on ASSP* 31(3).713–726.

RUSKE, G. 1986. Experiments on the use of demisyllables for automatic speech recognition. In *Montreal Symposium on Speech Recognition*, 49–50. McGill University.

RUSKE, G., & W. WEIGEL. 1992. Syllable-based stochastic models for continuous speech recognition. In *Speech Recognition and Understanding*, ed. by Pietro Laface & Renato De Mori, volume F 75 of *Nato ASI Series*, 193–198. Springer-Verlag.

SAGISAKA, Y., N. CAMPBELL, & N. HIGUCHI (eds.) 1996. *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer-Verlag.

SCHIEL, F., A. KIPP, & H. G. TILLMANN. 1998. Statistical modelling of pronunciation: It's not the model, it's the data. In (Strik *et al.* 1998), 131–136.

SELKIRK, E. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, Massachusettes: MIT Press.

SILVA, DAVID J. 1992. *The Phonetics and Phonology of Stop Lenition in Korean*. Ithaca: Cornell University.

SILVERMAN, KIM E. A. 1986. F0 segmental cues depend on intoantion: The case of the rise after voiced stops. *Phonetica* 43.76–91.

SILVERMAN, KIM E. A., 1987. *The structure and processing of fundamental frequency contours*. University of Cambridge dissertation.

SKALIČKOVÁ, ALENA. 1959. Some problems of general phonetics (Demonstrated on the system of Korean consonants). *Philologica* 1.29–39.

STRIK, H., & C. CUCCHIARINI. 1998. Modeling pronunciation variation for asr: overview and comparison of methods. In *Proceedings of the ESCA workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition'*, 137–144.

STRIK, H., J. M. KESSENS, & M. WESTER (eds.) 1998. *Modeling pronunciation variation for automatic speech recognition*, Rolduc, The Nethelands. European Speech Communication Association, University of Nijemegen.

'T HART, J., R. COLLIER, & A. COHEN. 1990. *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge University Press.

TALKIN, D. 1995. A robust algorithm for pitch tracking (RAPT). In *Speech Coding and Synthesis*, ed. by K. K. Paliwal. Elsevier.

TAYLOR, PAUL. 2000. Analysis and synthesis of intonation using the tilt model. *The Jounal of the Acoustical Society of America* 107(3).1697–1714.

TAYLOR, PAUL, & ALAN W. BLACK. 1994. Synthesizing conversational intonation from a linguistically rich input. In *2nd ESCA/IEEE Workshop on Speech Synthesis*.

TAYLOR, PAUL, HIROSHI SHIMODAIRA, STEPHEN ISARD, SIMON KING, & JAQUELINE KOWTKO. 1996. Using prosodic information to constrain language models for spoken dialogue. In *International Conference on Spoken Language Processing 96*, Rhode, Greece.

TERKEN, JACQUES. 1995. The perceptual relevance of micro-intonation: Enhancing the voicing distinction in synthetic speech by means of consonantal F0 perturbation. In *Studies in Applied Linguistics*, ed. by L. Hunyadi, M. Gósy, & G. Olaszy, volume 2, 103–124. Department of General and Applied Linguistics, Lajos Kossuth University of Debrecen, Hungary.

THORSEN, NINA. 1983. Two issues in the prosody of standard Danish. In *Prosody: Models and Measurments*, ed. by A. Cutler & D. R. Ladd, chapter 3, 27–38. Springer-Verlag.

UMEDA, N. 1981. Influence of segmental factors on fundamental frequency in fluent speech. *The Jounal of the Acoustical Society of America* 70.350–355.

USHER, M. J. 1984. *Information Theory for Information Technologists*. Macmillan.

WHALEN, D. H., & A. G. LEVITT. 1995. The universality of intrinsic f0 of vowels. *Journal of Phonetics* 23.349–366.

WIGHTMAN, C. W., & M. OSTENDORF. 1991. Automatic recognition of prosodic phrases. In *International Conference on Acoustics, Speech, and Signal Processing 91*, volume 1, 321–324, Toronto, Canada.

WILLEMS, NICO. 1982. *English intonation from a Dutch point of view*. Number 1 in Netherlands phonetic archives. Dordrecht, Holland: Foris.

WRIGHT, HELEN F., 2000. *Modelling Prosodic and Dialogue Information for Automatic Speech Recognition*. University Edinburgh dissertation.

YOSHIDA, KAZUNAGA, TAKAO WATANABE, & SHINJI KOGA. 1989. Large vocabulary word recognition based on demi-syllable Hidden Markov Model using small amount of training data. In *International Conference on Acoustics, Speech, and Signal Processing 89*, volume 1, 1–4.

YOUNG, S., J. JANSEN, J. OLLASON, & P. WOODLAND. 1996. *HTK Book*. Entropic.

YUN, SEONG JIN, HWAN JIN CHOI, & YUNG HWAN OH. 1997. Stochastic pronunciation lexicon modeling for large vocabulary continuous speech recognition. *The Jounal of the Acoustical Society of Korea* 16(2).49–57. (in Korean).

YUN, WEONHEE, & TAE-YEOUB JANG. 1999. Statistical analysis of Korean stop durations and its application for speech recognition. In *International Conference on Speech Processing 99*, volume 1, 305–310, Seoul, Korea.

ZHI, M. 1993. The duration of sound. *Sae-kuk-eo-saeng-hwal* 3(1).39–57. (in Korean).

ZHI, M., Y. J. LEE, & H. B. LEE. 1990. Temporal structure of Korean plosives in /VCV/. In *Proceedings of The Seoul International Conference on Natural Language*

Processing 90, 369–374, Seoul, Korea. Language Research Institute, Seoul National University.

APPENDIX A

Transcription of Phones

To transcribe pronunciation of Korean speech, I use only the characters specified in ASCII table, simply for the purpose of easy processing in machines. The list of those phones and corresponding IPA symbols are given in Table A.1. If an IPA symbol is ASCII compatible, it is used without modification. Three symbols [ŋ], [ɯ] and [ʌ] which cannot be represented by ASCII characters are replaced by [ng], [x], and [v] respectively. Note that the last two are vowels.

All relevant components of speech recognisers, such as lexicon, phone network, or unit list, are designed to use these machine readable symbols.

Identity	My Transcription	IPA Symbol
bilabial lax stop	p	p
bilabial tense stop	p'	p'
bilabial aspirated stop	p^h	p^h
alveolar lax stop	t	t
alveolar tense stop	t'	t'
alveolar aspirated stop	t^h	t^h
velar lax stop	k	k
velar tense stop	k'	k'
velar aspirated stop	k^h	k^h
palatal lax affricate	c	c
palatal tense affricate	c'	c'
palatal aspirated affricate	c^h	c^h
alveolar lax fricative	s	s
alveolar tense fricative	s'	s'
glottal fricative	h	h
bilabial nasal	m	m
alveolar nasal	n	n
velar nasal	ng	ŋ
lateral	l	l
flap	r	r
high front vowel	i	i
mid front vowel	e	e
high back rounded vowel	u	u
mid back rounded vowel	o	o
high back unrounded vowel	x	ɯ
mid central vowel	v	ʌ
low vowel	a	a
palatal glide	y	y
velar glide	w	w

Table A.1: Machine friendly phonetic transcription.

APPENDIX B

Romanisation of Korean

Romanisation of Korean characters is used for two purposes. First, all the processes dealing with texts with respect to recognition experiments is performed after the Romanisation (eg., lexicon construction, letter-to-sound rule, word list). This enhances the speed of processing considerably. Second, Korean text examples in the thesis are transcribed via Romanised form for international readability. Though the relevant pronunciation will be provided for better phonetic interpretation, other linguistic information needs to be expressed in terms of the orthographic system.

B.1 Conversion Table

Establishing criteria for transliteration from Korean characters into Roman characters is not simple because there is no direct one-to-one correlation between them. Among various versions of conversion methods, I followed the Romanisation standard agreed between the South and North Korean authorities in 1992. The only exception is for the conversion of the Korean liquid consonant which I have expressed consistently as 'l' rather than 'r' for onset and 'l' for coda as in the original mapping suggestion. I did not distinguish because the two sounds are a single phoneme at the abstract phonological

level. As Romanisation is just a character to character conversion, it does not need to describe phonetic realisation. The conversion map is in Table B.1. To prevent ambiguities, a boundary notation '-' is inserted between syllables.

It should be indicated that the Romanised characters are not to be confused with pronunciation symbols, as there is no one-to-one correspondence between them. For instance, multiple character symbols may share only one pronunciation symbol (ae, e $--> /e/$), while one character symbol can have multiple pronunciation (ui $-->$ [i] or [e]). Therefore, the number of consonants and vowels are not in accordance at two different levels of transcription.

B.2 Automation

Though the Korean writing system uses a combination of individual vowel and consonant characters, a set of ready combined syllables is most frequently used in computers. As each of them is represented by 8-bit 2-byte binary numbers, the key problem of Romanisation is to convert each syllable character into one or more 7-bit 1-byte English phone characters. Decomposing a syllable character into three syllable constituents (ie., onset, nucleus, and coda), and subsequently mapping each constituent to a Roman alphabet is a convenient way to do this.

As a syllable character consists of 16 bits (8-bit x 2-byte), each five bits from right to left can be assigned to coda, nucleus, and onset, respectively. The remaining left most bit is assigned to confirm whether the character is Korean or English. Then each of the 5-bit binary numbers is used, after converting it to a hexadecimal number, as a reference for corresponding to a relevant English alphabet. Figure B.1 is a summary of this procedure:

Korean	Roman	Example
ㄱ	k	*kuk* 'soup'
ㄴ	n	*nun* 'snow'
ㄷ	t	*ton* 'money'
ㄹ	l	*pal* 'foot'
ㄹ	r	*tali* 'leg'
ㅁ	m	*mom* 'body'
ㅂ	p	*pap* 'rice'
ㅅ	s	*son* 'hand', *os* 'clothes'
ㅇ	ng	*pang* 'room'
ㅈ	c	*cam* 'sleep'
ㅊ	ch	*chaek* 'book'
ㅋ	kh	*khal* 'knife'
ㅌ	th	*thal* 'masque'
ㅍ	ph	*phul* 'grass'
ㅎ	h	*him* 'strength'
ㅏ	a	*an* 'inside'
ㅔ	e	*te-ta* 'touch'
ㅣ	i	*i-mi* 'already'
ㅗ	o	*kho* 'nose'
ㅜ	u	*uri* 'we'
ㅐ	ae	*sae* 'bird'
ㅡ	eu	*kheun* 'big'
ㅓ	eo	*peol-le* 'bug'
ㅢ	ui	*ui-sa* 'doctor'
ㅘ	wa	*hwan-yeong* 'welcome'
ㅝ	weo	*kweon-lyeok* 'power'
ㅞ	we	*hwe-sa* 'company'
ㅙ	wae	*wae* 'why'
ㅑ	ya	*ya-ku* 'baseball'
ㅖ	ye	*ye-sul* 'art'
ㅛ	yo	*yo-sul* 'magig'
ㅠ	yu	*yu-lyeong* 'ghost'
ㅒ	yae	*yae-ki* 'story'

Table B.1: Korean-Roman character conversion table.

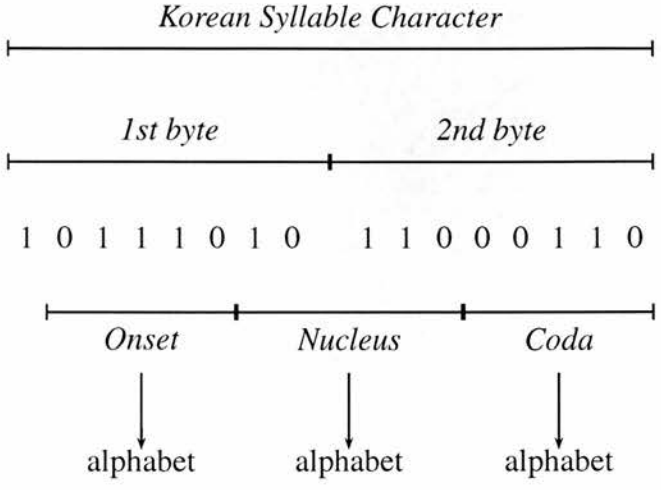


Figure B.1: Romanisation procedure. The leftmost bit of the first byte is used to denote the character is Korean.