Comparison of Phone Boundary Alignment between Handlabels and Autolabels*

Tae-Yeoub Jang** · Hyunsong Chung***

ABSTRACT

This study attempts to verify the reliability of automatically generated segment labels as compared to those obtained by conventional labelling by hand. First of all, an autolabeller is constructed using the standard HMM speech recognition technique. For evaluation, we compare the automatically generated labels with manually annotated labels for the same speech data. The comparison is performed by calculating the temporal difference between an autolabel boundary and its corresponding hand label boundary. When the mismatched duration between two labels falls within 10 msec, we consider the autolabel as correct. The results suggest that overall 78% of autolabels are correctly obtained. It is found that the boundary of obstruents is better aligned than that of sonorants and vowels. In case of stop sound classes, strong stops in manner-of-articulation wise and velar stops in place-of-articulation wise show better performance in boundary alignment. The result suggests that more phone-specific consideration is necessary to improve autosegmentation performance.

Keywords: Autosegmentation, Autolabelling, Speech Recognition, Time Alignment, Automatic Phone Recognition

1. Introduction

In many areas of spoken language processing and phonetic/linguistic experiments, the time domain phonetic annotation is a necessary procedure. Annotated data is a great asset both for phonetic research and for speech technologies. Though manual annotation

---

* This paper is the modified version of Jang & Chung (2002), which was presented at the 1st International Conference on Speech Sciences (May 10-11, 2002), Korea University, Seoul, Korea.

** Dept. of English, Hankuk University of Foreign Studies

*** Dept. of English Language Education, Daegu University

has been thought to be the most reliable way for these research areas, it has two major problems. First, it is quite slow and requires enormous human effort. Secondly, it is prone to inconsistent human errors. The first problem seems to have been effectively resolved by adopting the automatic labelling technique, a by-product of automatic speech recognition systems. The second problem also appears to be alleviated when we use automatic labellers. Although even the state-of-the-art automatic recognisers themselves still leave large room for improvement, producing a considerable amount of alignment errors in phone segmentation, at least the pattern of those errors are consistent, unlike human mistakes. One of the prominent advantages of autolabelling is its speed. Schiel *et al.* (1998) points out that more than two hours were spent in hand labelling of a 10 sec spontaneous utterance, for which an autolabeller would need no more than a minute. In regard to areas such as automatic speech recognition (ASR) and speech synthesis which rely on segmented speech corpora for training or construction purposes, fast automatic segmentation and labelling is critical for saving the overall construction time and simultaneously obtaining statistical stability. For more elaborate phonetic investigation, however, it is still uncertain whether these autolabels are accurate enough to carry reliable information on the basis of which rather subtle phonetic processes can be identified.

The main purpose of the current study is to investigate the extent to which a speech recogniser based autolabeller produces accurate phone labels compared with corresponding hand labels. Instead of intuitive, conceptual human judgment in terms of eye examination, an automatic method of comparing temporal information is used. For data, we used a single speaker (male) corpus of modern standard Korean, which was originally designed for statistical modelling of prosodic parameters for the construction of a speech synthesis system (Chung, 2002). In building an autolabeller, we employ the Hidden Markov Model (HMM) based statistical speech recognition techniques, as they, in general, exhibit superior time-alignment accuracy compared with earlier approaches, while providing a simple framework for decoding the phonetic boundary locations (Kemp *et al.*, 2000).

Although there are various previous attempts to improve the performance of the autolabeller (Vorstermans *et al.*, 1997; Jeong & Jeong, 1997; Svendsen & Kvale, 1990; Houben, 1989), the systematic studies, in which temporal alignment of handlabels and their corresponding autolabels is compared based upon different types of phones, are rare.

2. Data

The corpus used in this research was originally designed for the construction of a speech synthesis system. As the detailed description of the corpus construction is found in Chung (2002), only a brief summary is given here. First, a recording script was created. It has 670 sentence tokens extracted from news broadcast of Korea's two major broadcasting companies. It was the intention to use more natural material without using carrier-phrase style reading. The sentences had various lengths, so it was possible to analyse sentences with more than one prosodic phrase. Among them, 668 tokens are employed for the current analysis, as serious recording mistakes have been found in the two discarded files. The number of total phones in the corpus is 53,387. As is usually done for speech synthesis system constructions, a single subject was used; in this case, a 20-year-old male speaker of Korean. The recording was carried out in 12 sessions over a two-month time span. Though fewer sessions would have been ideal, it was physically impossible for a speaker to maintain the voice quality over 30 minutes of recording per day. The sentences were read as rapidly and fluently as possible to simulate a real news reading style. The recordings were made in an anechoic chamber on digital tape with sampling frequency, 16 kHz.

As the script is not artificially invented, no control of individual phone frequency is intended. As a consequence of this, there is a great difference in the number of tokens between phone classes. In other words, some phones occur more often than others, as will be clear in the later sections, but no effort was made to even out the distribution of phones. Thus the original distribution simply reflects how Korean phones are naturally distributed.

3. Labelling

The procedure of how two versions of label files were constructed is described in this section. Although constructing two versions of a file completely independent of each other would make the best comparison possible, we took the step of producing autolabel files first and hand-correcting them to generate the files of the hand label version. Nevertheless, the human labeller has tried not to be subject to previous information of boundary location given by the autolabeller throughout the labelling works.

3.1 Phone Sets

We adopt the 38 phone units of Korean for segmentation and annotation of sentences, as follows:

Obstruents (15): c (palatal affricate) c'(glottalised) ch (aspirated) k k' kh p p' ph t t' th h s s'

Sonorants (5): m n ng l r

Monophthongs (7): a e i o u v (mid central vowel) x (high back unrounded vowel)

Diphthongs (10): wa we wi wv xi ya ye yo yu yv

Silence (1): sil

All machine readable symbols, instead of other phonetic symbols, are used but they can be easily converted to corresponding phonetic symbols such as International Phonetic Alphabet (IPA) (see Jang (2000), for such conversion). Note that glides are not treated as separate phones but only as a part of diphthongs. Accordingly, each diphthong is regarded as a single unit rather than a mixture of a vowel and a glide. This is for the convenience of the automatic phone recogniser construction which will be described later in this paper. The symbol ' sil' is used only for utterance initial silence demarcation and all the other sentence internal silence tokens, such as *pause*, appearing next to utterance internal intonational phrase (IP) boundaries, are not counted in the current study.

3.2 Autolabelling

Autolabels are produced taking advantage of an automatic speech recognition technique. The core of autolabelling process is constructing a phone-level speech recogniser. As for the phone-level speech recogniser, the recognition target is the phone string rather than the word string as is the case for normal automatic speech recognition systems. It is not appropriate to use the current data for constructing a phone recogniser for two reasons. Firstly, as already mentioned, the corpus is made for speech synthesisers, so it is not large enough for training a recogniser. Secondly, and more crucially, we are not supposed to use the same data for constructing and testing the same autolabeller. Doing so is just as illegible as using the same data for training and testing a single speech recogniser.

Subsequently, we used a phone recogniser built completely independent of the targeted experiment described in this paper. The database used to build the recogniser is also a different one known as a KAIST database (Park *et al*, 1995), which is composed of about 10,000 sentence tokens spoken by more than 100 speakers. For training phone models, a well-known statistical approach, Hidden Markov Model (HMM) technique was used. More specifically, a 3-state left-to-right continuous HMM was established for each phone in terms of acoustic feature parameter estimations along with a statistical bigram language(phone) model for constraining sequence of phones. The more detailed description of this procedure is shown in Jang (2000). During the course of the entire processing, a useful tool called the Hidden Markov Model Toolkit (HTK v3.0: Young *et al*, 1996) was used. Once the phone recogniser is constructed we can directly use it as an autolabeller without any major modification. Phone sequence language models, orthographic transcription, and a lexicon of word-to-phone mapping were created in advance and they were provided when the autolabeller was working on our target data tokens. This procedure is briefly illustrated in Figure 1. All 668 tokens were autolabelled in this way.

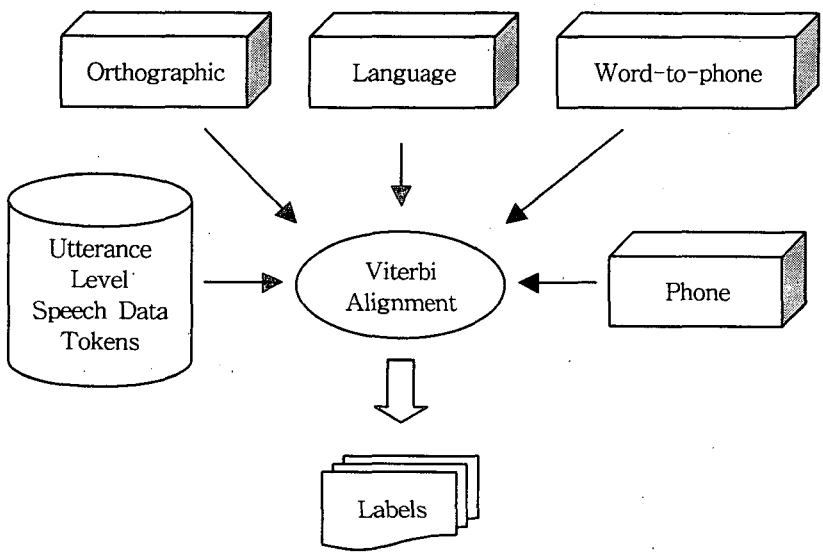

Figure 1. Automatic labelling procedure

3.3 Hand Labelling

We need another version of label files which are to be compared with corresponding autolabels. Instead of starting from scratch, we utilise the autolabels for creating hand

labels. An autolabel file for each sentence, together with its corresponding waveform and spectrogram, is opened on a speech analysis tool known as *xwaves* (Entropic, 1998). All the individual phone labels are closely checked and the demarcation boundaries are shifted to and fro whenever the human labeller finds it necessary. Deletions and insertions of phones are also possible on the basis of the labeller's observation and judgment. For the consistency of work, only a single human labeller (one of the co-authors of this paper) took charge of labelling all the tokens. Detailed labelling criteria can be described as follows:

- Stops and affricates were annotated with information of the closure duration, the burst and aspiration. In post-pausal position, 50 msec of closure duration was arbitrarily included. The stop closure onset in pre-pausal position was defined as the point at which energy in the region of F2 and the higher formants ceases to be visible on the spectrogram display. The closure duration of the stop in pre-pausal position was assumed to be 20 msec. Aspiration was marked as the duration between the burst and the first glottal pulse in the vowel. When stops or affricates were preceded by another stop, 30 msec of closure duration before the consonants were arbitrarily allocated to the second consonant and the rest closure duration was assigned to the preceding stop. When stops preceded fricatives, they were annotated from the end point of the stop to the point when a change of zero-crossing rate was found between two consonants.
- Fricatives were annotated when high-frequency energy appeared. When it was difficult to find high-frequency energy, a change in zero-crossing rate was used as a cue to define the onset or the offset of fricatives.
- Nasal boundaries were defined as the points when formant frequencies showed a discontinuity and the amplitudes of the formants were decreased. When two nasals were geminated, their boundary was determined from any change in the energy, otherwise the mid-point was chosen.
- In the lateral [l], in some cases the amplitude of F1 was decreased. The mid-point of this transition was assumed to be the onset of the lateral [l]. The flap [r] was easily detected, because it only appeared between vowels. When it appeared, it had a 20 - 30 msec duration with an energy decrease and weaker formants.
- Both diphthongs and monophthongs were considered unitary vowels. Whenever formant frequencies appeared after consonants, they were considered to indicate

the vowel onset. A diphthong was treated as a glide and a monophthong. An energy change could be observed in the onset of glides, but in this experiment, no boundary was annotated between the glide and the rest of the vowel. Nasalised vowels were annotated as oral vowels. In nasalised vowels, the amplitude of F2 and F3 were decreased. When two vowels were adjacent to each other, the formant change was first investigated. If there was still a difficulty in distinguishing the boundaries, the energy change between two vowels was used. Otherwise, the boundary was annotated at the mid-point.

Although the hand label creation is based upon autolabels, and consequently this may be one of the points which can be more or less criticised, it is assumed that the elaborate eye examination has sufficiently abolished the effect of autolabel basis.

4. Procedure of Automatic Comparison

4.1 Adjustment of Label Files

The phone string of a hand label file is not supposed to be identical to that of the corresponding autolabel file, since there have been frequent corrections including insertion or deletion of phones in the case of hand labels. Thus, the adjustment of phone strings is necessary for two label files to be compared in parallel. Those inserted and deleted phones are not taken into account since they do not have a comparable counterpart in the corresponding label file. In the case of substitution of phones, the phone in the hand label is thought to be the right one and the corresponding phone in the autolabel file is corrected as such. For example, when, as is often the case, an autolabel file contains [e] while the corresponding hand label file has [xi] in its place, [xi] is replaced by the symbol [e]. All this procedure has also been automated.

4.2 Evaluation Methods

The duration between a hand label boundary and its corresponding autolabel boundary is calculated for each phone symbol. The temporal information of boundaries is the ending point of the relevant phone. For example, if an autolabelled [a] has the ending boundary at 35-msec point, while the corresponding hand label [a] has the ending boundary at 48-msec point, then their difference in time, namely, 13 msec is

acquired and preserved. A program is written to automatically calculate the difference for all the phones.

Although we will show and discuss the results according to various mismatched duration levels, the 10-msec point is critically regarded as a threshold and the misalignments longer than 10 msec are thought to be real errors. Henceforth, in this paper, we would represent such error as BAE (Boundary Alignment Error), which can be defined as "an alignment error calculated in terms of duration (msec) between an autolabel boundary and its corresponding handlabel boundary". Correspondingly, another representation, BAA (Boundary Alignment Accuracy), the rate of boundary label pairs which fall within the 10-msec range, is also used.

5. Results and Discussion

Table 1. Boundary mismatch errors: it means, for example, that 78.63 % of all the "hand/auto label" boundary mismatches are by less than 10 msec.

Phone	Number of Tokens	0-10 msec (%)	10-20 msec (%)	20-30 msec (%)	over 30 msec (%)
a	4713	70.21	20.45	5.47	3.86
c	1796	79.57	17.48	2.28	0.67
c'	238	81.51	17.65	0.84	0
ch	664	81.63	16.87	1.20	0.30
e	2738	69.14	21.69	5.00	4.16
h	1172	91.13	4.44	1.88	2.56
i	4536	79.56	14.22	3.46	2.76
k	3493	86.60	10.76	1.69	0.94
kh	318	86.79	10.38	2.20	0.63
k'	396	88.13	11.62	0.25	0
l	1710	80.99	12.75	2.98	3.27
m	2215	86.64	9.30	2.71	1.35
ng	1925	88.00	8.88	1.56	1.56
n	5494	81.67	12.41	2.17	3.75
o	2309	72.02	15.16	5.15	7.67
p	1470	76.26	17.28	4.69	1.77
ph	384	85.42	12.24	1.56	0.78
p'	80	80.00	18.75	0	1.25
r	1458	52.61	29.29	12.21	5.90
s	2106	85.28	13.25	0.95	0.52
sil	668	83.68	8.98	1.80	5.54
s'	762	94.36	3.94	0.79	0.92

Table 1. continued

t	2465	73.31	21.05	4.26	1.38
th	386	83.68	14.77	1.55	0
t'	335	88.06	10.75	0.30	0.90
u	1512	83.07	11.38	3.44	2.12
v	2249	69.99	20.54	6.18	3.29
wa	394	57.36	27.66	9.14	5.84
we	358	72.35	20.67	4.47	2.51
wi	131	73.28	13.74	3.05	9.92
wv	181	69.61	20.44	4.97	4.97
xi	62	88.71	6.45	1.61	3.23
x	2861	86.96	9.44	2.03	1.57
ya	112	79.46	9.82	6.25	4.46
ye	111	76.58	16.32	3.60	4.50
yo	246	68.29	19.51	6.10	6.10
yu	242	74.38	19.83	2.89	2.89
yv	1097	68.73	19.96	7.20	4.10
Total	53387	78.63	15.02	3.56	2.79

Table 1 shows phone specific results of comparison between two versions of labels. The overall result shows that 78.63% (41,977 pairs) of all the compared pairs (53,387) are within the difference range of "below 10 msec". The duration "10 msec" is meaningful in that our automatic segmentation program also has the 10-msec duration of frame shifting. Thus, in the digital speech processing point of view, counting duration information less than 10 msec is just pointless. Besides, as a typical male voice is produced with roughly 10-msec-long glottal pulses, the duration shorter than this cannot be considered to contain voice-characterising information. Therefore, the alignment errors of which the size is below 10 msec may not be regarded as significant, and only the errors larger than 10 msec should be considered seriously and counted as true errors.[1] In the table, 22.37% of all the alignment cases are shown to be such errors (BAE), which is still a considerable amount. If we regard misalignment up to 20 msec as tolerable, the BAE reduces down to only 6.35%. Here, it should be noted that these errors are not necessarily caused by automatic labelling alone. Even if the hand labels

1) For Korean stop consonants, this 10 msec limit can cause a problem as their manners of articulation are very frequently distinguished and characterized in terms of sensitive temporal information such as VOT (voice onset time). This problem needs to be further investigated in future studies.

are verified to be free from major mistakes, a certain degree of inconsistency of human labellers are inevitable and it also seems responsible for some alignment mismatches.

Table 2. Mismatch errors for each phone class

Phone class	Number of Tokens	0-10 msec (%)	10-20 msec (%)	20-30 msec (%)	over 30 msec (%)
V (monophthongs)	20918	75.49	16.53	4.40	3.58
V (diphthongs)	2934	69.46	19.94	6.07	4.53
C (non-continuants)	12025	81.11	15.39	2.54	0.96
C (continuants)	4040	88.69	8.94	1.19	1.19
Nasals	9634	84.08	10.99	2.17	2.76
Liquids	3168	67.93	20.36	7.23	4.48

Table 2. shows results calculated upon each natural class of phones. It is shown that liquids are worst in accuracy. But this result is mainly because of the flaps whose accuracy (BAA) is as low as 52.61, as given in Table 1, barely above the half point. This is understandable when we consider the acoustic characteristics of flap sounds. It has been reported that they have considerably lower energy and shorter duration than other phones (Song *et al.*, 1995). As it appears to be weak and short at the display of speech analysis tools, human labellers tend to locate their ending points a little behind the true ending points making their durations look longer than what they really are. Our data supports this inference, as it was found that the overall average duration of flap sounds in the autolabels was shorter (73.10 msec) than those in hand labels (91.70 msec)

Generally speaking, apart from liquids (flaps, more specifically), obstruents appear to be more accurate than vowels. Especially diphthongs turn out to have more errors than monophthongs. The relatively long duration of diphthongs appears to be one of the reasons for such mismatches.

Considering that the distribution of the Korean consonantal system the which is dominated by non-continuants such as stops and affricates, and that those obstruents are frequently the cause of human manual phonetic annotation, it is worth investigating and comparing the accuracy of stop sounds more in detail. The accuracy for each obstruent class is revealed in Table 3 and compared in Figure 2 and 3. It appears that there is no critical difference in accuracy between affricates and stops. When it comes to place-of-articulation specific stop internal comparison, it is obvious that velar stops [k, kk, kh] are more accurate than the other two types, although the reason for this bias is

not clear at present. As for the manners of articulation, it is shown that the boundaries of strong (tense & aspirated) stops are better aligned than weak (lax) ones. This tendency is also found in affricates (see Table 1) although it is not as prominent as in stops. This result seems to have much to do with the fact that lax stops are far more susceptible to phonological processes and very frequently lose their original identity resulting in voiced sounds or even approximants. Thus, it seems hard to detect the boundaries of these sounds whether manually or automatically, consequently causing less consistency in alignment. In a technical point of view the apparent existence of a silence period at the strong stops helps both men and machine find the boundary between the stop sounds and adjacent sounds.

Table 3. Boundary Alignment Accuracy (BAA; less than 10 msec mismatch) of non-continuants.

Phone class		Number of Tokens	BAA (%)
Affricates		2,698	80.24
Stops		9,327	81.36
Stops (Place specific)	Bilabial	1,934	78.23
	Alveolar	3,186	76.11
	Velar	4,207	86.76
Stops (Manner specific)	Lax	7,428	80.14
	Tense	811	87.30
	Aspirated	1,088	85.20

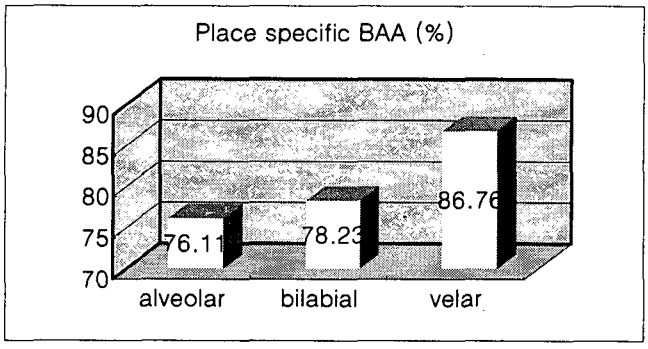

Figure 2. BAA comparison among stop places

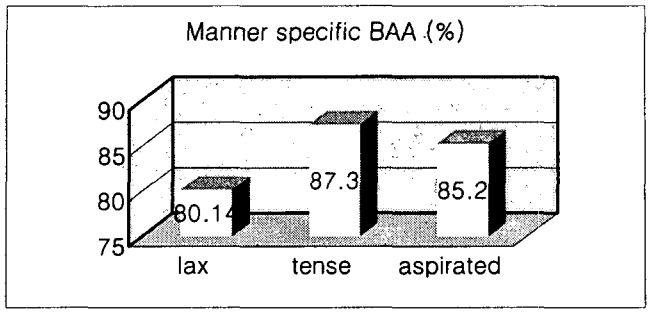

Figure 3. BAA comparison among stop manners

6. Concluding Remarks

What we mainly found in this study is that automatically generated phone labels considerably agree with the corresponding manual labels in terms of boundary location information. Especially, the performance is relatively better for obstruents and nasals, which suggests that methods of improving autolabelling performance for the phones of other classes need to be sought in order to achieve more reliable autolabels. We also show that among stop sound classes, velar stops are more accurately aligned than stops produced at the other places of articulation whereas tense and aspirated stops are better aligned than lax stops.

Other than stops, it is not clear, at present, how significant the difference in performance depending upon various phone classes is. Nor can we expect similar results from experiments using other language data. Further experiments with other corpora of various languages as well as of Korean will clarify these questions.

References

Chung, H. 2002. *Analysis of the Timing of Spoken Korean with Application to Speech Synthesis*. Ph.D. Thesis, University College, University of London.

Entropic. 1998. *The Manual of ESPS/Waves+ with EnSig5.3. Version 5.3.*

Houben. C. G. J. 1996. "Knowledge based parameters for HMM speech recognition." *Proceedings of ICASSP 1996*, 29-32.

Jang, T. Y. 2000. *Phonetics of Segmental F0 and Machine Recognition of Korean Speech*. Ph.D. Thesis. University of Edinburgh.

Jang, T. Y. & H. Chung. 2002. "How reliable are autolabels?" *Proceedings of the 1st International Conference on Speech Sciences, Addendum.* The Korean Association of Speech Sciences, 20-26.

Jeong, C. G. & H. Jeong. 1997. "Automatic phone segmentation and labelling of continuous speech." *Speech Communication*, 20, 291-311.

Kemp, T., M. Schmidt, M. Westphal & A. Waibel. 2000. "Strategies for automatic segmentation of audio data." *Proceedings of ICASSP 2000*, 1423-1426.

Park, J. R., O. W. Kwon, D. Y. Kim, I. J. Choi, H. Y. Jeong & C. K. Un. 1995. "Speech data collection for Korean speech recognition." *The Journal of the Acoustic Society of Korea*, 14(4), 74-81.

Schiel, F., A. Kipp & H. G. Tillmann. 1998. "Statistical modelling of pronunciation: It's not the model, it's the data." In H. Strik, J. M. Kessens and M. Wester (eds.), *Modeling Pronunciation Variation for Automatic Speech Recognition*, 131-136.

Song, M., T. Y. Jang & H. Chung. 1995. "The acoustic properties of the flap in natural sentence utterance." *Proceedings of the 1995 Spring Conference on Cognitive Science*, Korea Cognitive Science Society (in Korean), 11-17.

Svendsen, T. & K. Kvale. 1990. "Automatic alignment of phonetic labels with continuous speech." *Proceedings of ICSLP 1990*, 997-1000.

Vorstermans, A., J. P. Martens & B. Van Coile. 1997. "Automatic segmentation and labelling of multi-lingual speech data." *Speech Communication*, 19, 271-293.

Received: January 30, 2003
Accepted: February 28, 2003

▲ Tae-Yeoub Jang
Department of English, Hankuk University of Foreign Studies
270, Imun-dong, Dongdaemun-gu, Seoul, 130-791, Korea
Tel: +82-2-961-4770 (O) +82-2-535-3637 (H)
Fax: +82-2-965-2183
E-mail: tae@hufs.ac.kr

▲ Hyunsong Chung
Department of English Language Education, Daegu University
15, Naeri, Jinryang, Gyeongsan, Gyeongbuk, 712-714, Korea
Tel: +82-53-850-4125 (O) +82-53-215-1739 (H)
Fax: +82-53-850-4121
E-mail: hchung@daegu.ac.kr