THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.24, NO.2E 2005. 6 pp. 71~76

# An Analysis of Acoustic Features Caused by Articulatory Changes for Korean Distant-Talking Speech

Sunhee Kim*, Soyoung Park*, Chang D. Yoo*
*Dept. of Electrical Engineering and Computer Science, KAIST
(Received May 24 2005; revised May 30 2005; accepted Jun 8 2005)

**Abstract**

Compared to normal speech, distant-talking speech is characterized by the acoustic effect due to interfering sound and echoes as well as articulatory changes resulting from the speaker's effort to be more intelligible. In this paper, the acoustic features for distant-talking speech due to the articulatory changes will be analyzed and compared with those of the Lombard effect. In order to examine the effect of different distances and articulatory changes, speech recognition experiments were conducted for normal speech as well as distant-talking speech at different distances using HTK. The speech data used in this study consist of 4500 distant-talking utterances and 4500 normal utterances of 90 speakers (56 males and 34 females). Acoustic features selected for the analysis were duration, formants (F1 and F2), fundamental frequency, total energy and energy distribution. The results show that the acoustic-phonetic features for distant-talking speech correspond mostly to those of Lombard speech, in that the main resulting acoustic changes between normal and distant-talking speech are the increase in vowel duration, the shift in first and second formant, the increase in fundamental frequency, the increase in total energy and the shift in energy from low frequency band to middle or high bands.

***Keywords:*** *Distant-Talking Speech, Normal Speech, Lombard Effect, Acoustic Features, Articulatory Changes*

## I. Introduction

When a speaker controls a recognition system located at a certain distance, the use of close-talking microphones is not feasible[1]. And the speaker in this distant-talking environment tends to increase his/her vocal level to avoid recognition failure. This effort by the speaker to speak more clearly increases intelligibility for human-to-human communication, while it causes degradation of the recognition system[2], along with the distance effect such as interfering sound and echoes. In other words, compared to close-talking speech, the distant-talking speech degrades the performance of speech recognizer, as the acoustic data captured with a microphone from some distance to the speaker are corrupted, on the one hand, by the distant effect, and on the other hand, by the articulatory variability made by speakers to improve intelligibility.

Within speech recognition technology, the user's vocal effort which causes degradation of the system is also found in noisy environments, which is so-called the Lombard effect, a phenomenon caused by the articulatory changes made by speakers in order to be more intelligible in the noisy environment[3]. According to previous works on the acoustic-phonetic characteristics of the Lombard effect[3, 4, 5, 6], many acoustic and perceptual characteristics depend on the speakers. Despite the differences such as the recording conditions and the number and

Corresponding author: Sunhee Kim (shkim@ee.kaist.ac.kr)
EECS Dept. KAIST 373-1 Guseong-dong Yuseong-gu Daejeon. Korea

characteristics of speakers, similar tendencies are observed across languages. The main acoustic changes between normal speech and Lombard speech are: (i) increase in fundamental frequency (F0); (ii) shift in energy from low frequency bands to middle or high bands; (iii) increase in level; (iv) increase in vowel duration; (v) spectral tilting; and (vi) shift in formant center frequencies for F1(mainly) and F2[3]. General methods used to improve the performance of speech recognition algorithms for Lombard speech were summarized in three general areas[5]: (i) those of robust features; (ii) those of equalization methods, (iii) those of model adjustment or training methods.

Studies on Lombard speech of particular languages have been conducted for American English, French, Spanish and Japanese [3]. Lombard speech was found to be distinct from normal speech in similar ways for the four languages. In case of the Korean language, [7] proposed a Lombard effect compensation model simulating the variations of formant location, formant bandwidth, pitch, spectral tilt, and energy in each frequency band. [8] proposed a robust feature extraction method using Lombard effect compensation filter. Both studies focused on the compensation aspect only, and rarely has studies been conducted on acoustic features caused by articulatory changes for Korean either in Lombard speech or in distant-talking speech.

This paper will review the speech recognition experiments conducted for normal speech as well as distant-talking speech at different distances using HTK in order to show the effect of different distances and articulatory changes. In addition, the tendencies of some acoustic features caused by articulatory changes for Korean distant-talking speech in different phone classes with regard to normal speech will be examined and the influence of the gender of the speakers in the characterization of these tendencies will also be analyzed.

This paper is organized as follows: Section 2 describes the speech database used for this analysis and Section 3 presents the results of recognition experiments performed on normal speech and distant-talking speech. Section 4 describes the experiment framework, Section 5 the experimental results and discussion, and Section 6 concludes the paper.

## II. The speech database

The acoustic-phonetic analysis was performed on the speech data designed in [7], which consist of 18000 distant-talking utterances and 4500 normal utterances of 90 speakers (56 males and 34 females). Each speaker was made to pronounce the list twice in normal speech and three times in distant-talking speech. After our evaluation of the recorded utterances, one utterance for each style per speaker was saved.

The recording was conducted in a quiet room located on campus. In order to obtain normal utterances, each speaker was asked to pronounce in normal style each word appearing on the screen with a 3-second pause between two words. An unidirectional microphone (UD-252 by McCanon Inc.) was placed at a distance of 0.5 m from the speaker for the recording of normal speech. For the acquisition of distant-talking speech, a voice command controlled PC was placed at a distance of 3 m. And the utterances were recorded through the same kind of unidirectional microphones used for the recording of normal speech at a distance of 0.5 m, 1 m, 2 m, and 3 m simultaneously. Accordingly, the distant-talking speech at each distance was 4500 utterances respectively.

For the recording, 50 isolated words for PC command were pronounced in normal style and in distant-talking speech. They were composed of 4 one-syllable words, 36 two-syllable words, 6 three-syllable words and 4 four-syllable words. The speech database were sampled using a 16-bit A/D converter at a sample rate of 16 kHz.

In this study, the recordings captured by microphone at a distance of 0.5 m form the speaker (4500 Utterances), that is, at the same distance for normal speech from the speaker, were analyzed in order to compare the acoustic features caused by articulatory changes in distant-talking speech with those produced in normal speech (4500 Utterances).

## III. Distant-talking speech recognition

In order to examine the effect of the different distances and the articulatory changes, speech recognition experiments were conducted for normal speech as well as distant-talking speech at different distances using HTK. For the recognition experiment, the left-to-right HMM based on the triphone was used, and each model was organized by 5 states with 4 Gaussian mixtures. For the characteristic features, 39-dimensional MFCC were used.

For the experiment, only the speech databases recorded by 34 female speakers were used. Among 1700 normal utterances (50 words by 34 speakers), 1360 utterances were used for training

Table 1. Recognition rates of speech recognition experiments for normal speech and distant-talking speech.

| | Normal | Distant-talking | | | |
|---|---|---|---|---|---|
| | | 0.5m | 1m | 2m | 3m |
| Rec. rate (%) | 95.46 | 86.27 | 78.93 | 76.41 | 73.89 |

data and 340 for test data. The test for normal speech was performed 5 times using a random shuffle algorithm and the final result corresponds to the average recognition rate of 5 tests. The same training data were used for experiments of distant-talking speech at each distance (1700 utterances at each distance). Table 1 shows the recognition rates for normal speech and distant-talking speech of different distances.

As shown in Table 1, the performance of the recognizer degraded from 95.46% for normal speech to 73.89% for distant-talking speech at 3m away. As the distant-talking speech made at a distance of 0.5 m was captured in exactly the same condition as normal speech, it is possible to interpret that the recognition rate difference between these two speeches lies in articulatory changes. However, the recognition rate of distant-talking speeches made at distances between 1 and 3 m could be interpreted that they are influenced by both articulatory changes and distance effect such as interfering sound and echoes. In sum, the recognition rates are influenced not only by distance but also by articulatory variability made by speakers to improve intelligibility.

# IV. Experimental framework

Table 2. Korean consonants.

| | | labial | dental | palatal | velar | glottal |
|---|---|---|---|---|---|---|
| stop | lenis | p | t | | k | |
| | fortis | p' | t' | | k' | |
| | aspirated | ph | th | | kh | |
| affricate | lenis | | | c | | |
| | fortis | | | c' | | |
| | aspirated | | | ch | | |
| fricative | lenis | | s | | | h |
| | fortis | | s' | | | |
| nasal | | m | n | | | ŋ |
| liquid | | | l/r | | | |

Table 3. Korean vowels.

| | anterior | posterior | |
|---|---|---|---|
| | unround | unround | round |
| high | i | ɯ | u |
| middle | e | ʌ | o |
| low | | a | |

## 4.1. Korean phonetic units

Table 2 and Table 3 present Korean consonants and vowels transcribed in IPA (International Phonetic Alphabet).

As the word list for the speech data consists of 50 words for PC command, it is not completely phonetically balanced and

The 2 consonants ($p^i$ and $k^h$) which do not influence the results of the current analysis, were not included. The total number of phonetic units appearing in the word list were 241, which consisted of 41 stops, 17 fricatives, 28 affricates, 41 nasals, 19 liquids and 95 vowels. Semi-vowels (/w/ and /y/) were excluded from our analysis in this paper because of the risk of error that may result during their segmentation and labeling process due to their short duration and transitional characteristics.

## 4.2. Acoustic features

A large variety of acoustic features have been studied to analyze the Lombard effect with respect to normal speech[6, 9]. The acoustic features considered in this work are as follows.

- Phonetic Duration (PD)
- First Formant (F1)
- Second Formant (F2)
- Fundamental Frequency (F0)
- Total Energy (TE)
- Energy between 0-250Hz
- Energy between 250-500Hz
- Energy between 500-1000Hz
- Energy between 1000-2000Hz
- Energy between 2000-3000Hz
- Energy between 3000-4000Hz
- Energy between 4000-5000Hz
- Energy between 5000-6000Hz
- Energy between 6000-7000Hz

## 4.3. Procedure

As already mentioned above, the distant talking speech analyzed in this study is the speech captured at a distance of 0.5m. Each utterance was segmented and labeled by forced alignment provided in HTK. Programs for the extraction of acoustic features were developed using Matlab and C++. And the acoustic features of distant-talking speech in comparison with normal speech was statistically analyzed by phone class using SPSS 12.0. For each phone class, the changes of distant-talking speech compared with normal speech for selected acoustic

features were observed. The percentage variation of each feature is obtained by the following calculation: 100*(distant talking average-normal average)/normal average. A 0.05 level of significance was established to test whether the differences were significant between distant-talking and normal values corresponding to each feature for each phone class.

## V. Results and discussion

### 5.1. Results for phone duration, first and second formants, fundamental frequency and total energy

Table 4 shows the percentage variations of distant-talking speech compared with normal speech for each phone class. The values in the shaded blanks were not estimated as they are irrelevant for corresponding phone classes, and the values in shaded cells are those which do not have any statistical significance.

For all speakers, normal speech was statistically different from distant-talking speech with respect to all acoustic features of concern. The phone duration was shown to decrease for stops, fricatives, affricates and nasals, while it was shown to increase for liquids and vowels. Further, the amount of increase for liquids and vowels was much larger than that of the decrease for unvoiced groups. Similar results were reported for English Lombard speech, in which case consonants tended to decrease while vowels increased in duration[4]. For Spanish, only vowels showed a small increase with statistical significance[6]. For the formants, the shifts were observed for both first formant and second formant. The amount of the shift for the first formant was considerably larger than that for the second formant. The results mostly corresponded to those observed for Spanish[6]. In regard to the fundamental frequency and total energy, increases were observed in all phone classes, which corresponds to the results for English[4] and Spanish[6].

With respect to the percentage variation of acoustic features for each phone class depending on the gender of speakers, the phone duration did not show any regularity across whole classes. However, for some classes, male speakers showed a larger decrease or increase, and female speakers for other classes. In case of the first formant, the amount of change for female speakers was larger than that for male speakers except for liquid, while the amount of change for female speakers was larger than that for male speakers in all cases for the second formant. The tendency was inverse for fundamental frequency, and the amount of change was always larger in male speakers. In case of the total energy, the tendency did not seem to be affected by gender, so that, for each gender, it had similar values of change. The homogeneous difference between male and female speakers also was reported for fundamental frequency in Spanish [6], but not for other features in other languages.

Table 4. Percentage variations of acoustic features for each phone class for male speakers (M), female speakers (F) and for all the speakers together (all) (PD: phone duration, F1: first formant, F2: second formant, F0: fundamental frequency, TE: total energy).

| | | PD | F1 | F2 | F0 | TE |
|---|---|---|---|---|---|---|
| Stops | M | -3.2 | | | | 14.3 |
| | F | -12.8 | | | | 13.5 |
| | all | -6.8 | | | | 14.0 |
| Fricatives | M | -9.4 | | | | 16.4 |
| | F | -6.5 | | | | 15.3 |
| | all | -8.3 | | | | 16.0 |
| Affricates | M | -5.7 | | | | 18.9 |
| | F | -9.5 | | | | 19.4 |
| | all | -7.1 | | | | 19.1 |
| Nasals | M | 3.06 | 29.5 | 1.49 | 87.9 | 10.6 |
| | F | -12.3 | 53.5 | 4.6 | 50.1 | 5.2 |
| | all | -2.7 | 38.6 | 2.7 | 70.6 | 8.5 |
| Liquids | M | 18.5 | 32.1 | 1.48 | 64.8 | 16.9 |
| | F | 47.2 | 30.9 | 6.4 | 45.6 | 16.3 |
| | all | 29.4 | 31.6 | 3.3 | 42.9 | 16.6 |
| Vowels | M | 62.4 | 13.4 | 2.53 | 76.1 | 20.6 |
| | F | 86.9 | 26.6 | 3.25 | 55.9 | 22.2 |
| | all | 71.7 | 18.4 | 2.8 | 50.8 | 21.2 |

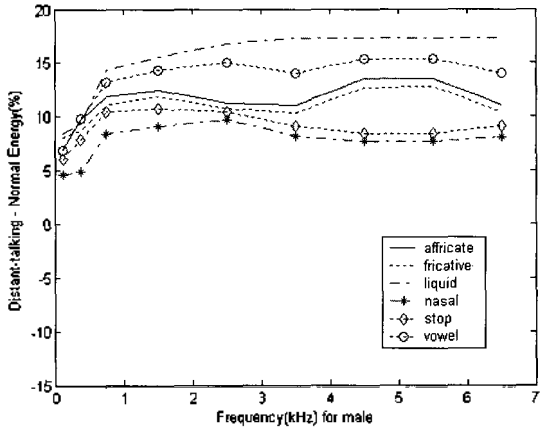


Figure 1. Percentage variations of the energy distribution for each phone class for male speakers.

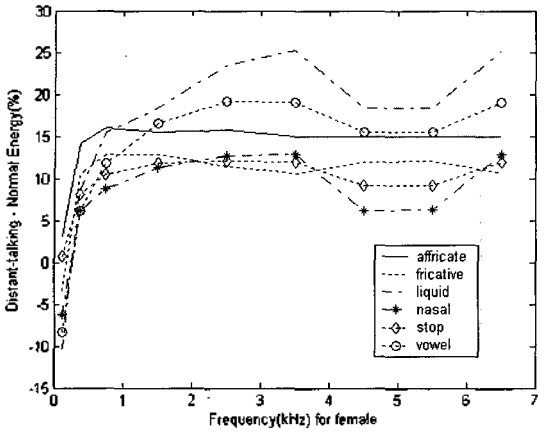


Figure 2. Percentage variations of the energy distribution for each phone class for female speakers.

## 5.2. Results for energy distribution

In Figure 1-2 , the percentage variations of the spectral energy distribution are shown for each phone class for male and female speakers respectively.

The shifts in the energy from low frequency bands to middle or high bands are observed for all cases. And the tendencies are almost the same for both female and male speakers. In low frequency bands, the energy decreases for female speakers but it increases for male speakers. Similar tendency was also reported in Spanish[6].

# VI. Conclusions

This paper analyzed the tendencies of some acoustic features in different phone classes for Korean distant-talking speech due to the articulatory changes with regard to normal speech. The influence of gender of the speakers in the characterization of these tendencies were also studied. In order to examine the effect of the different distances and the articulatory changes, speech recognition experiments were conducted for normal speech as well as distant-talking speech at different distances using HTK.

The speech recognition experiments confirmed that the recognition rates were influenced not only by distance but also by articulatory variability made by speakers to improve intelligibility. The main acoustic changes between normal and distant-talking speech are summarized as follows. The decrease in the duration was observed for stops, fricatives, affricates and nasals, while an increase was observed for liquids and vowels. Also, increases in the first and second formants were observed, and the amount of increase for the first formant was considerably larger than that for the second formant. In case of fundamental frequency and total energy, the increases were observed in all cases. In terms of energy distribution, energy was shifted from low frequency bands to middle or high bands. Similar to the results in Spanish, the influence of gender was found in fundamental frequency and energy distribution.

In conclusion, the results showed that the acoustic changes resulting from the speaker's effort to be more intelligible in distant-talking speech were similar to those produced in Lombard speech achieved by previous studies. Therefore, the study for performance improvement of distant-talking speech recognition could be performed with the help of the Lombard effect compensation methods proposed in earlier literature.

## Acknowledgements

This work was supported by Brain Korea 21 Project, the School of Information Technology, KAIST in 2005.

## References

1. M. Matassoni, M. Omologo, D. Giuliani, P. Svaizer, "Hidden Markov model training with contaminated speech material for distant-talking speech recognition" Computer Speech & Language, 16 (2), 205-223, 2002.
2. S. Köster, "Acoustic Characteristics of Hyperarticulated speech for Different Speaking Style," Proc. ICASSP, 2, 873-876, 2001.
3. J.-C. Junqua, "The Influence of Acoustics on Speech Production: A Noise-Induced Stress Phenomenon as the Lombard Reflex," Speech Communication, 20, 13-22, 1996.
4. J. Hansen, "Analysis and Compensation of Speech under Stress and Noise for Environmental Robustness in Speech Recognition," Speech Communication, 20, 151-173, 1996.
5. S. E. Bou-Ghazale, J. Hansen, "A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress, " IEEE Transactions on Speech and Audio Processing, 8-4, 429-442, 2000.
6. A. Castellanos, J.-M. Benedi, F. Casacuberta, "An Analysis of General Acoustic-Phonetic Features for Spanish Speech Produced with the Lombard Effect", Speech Communication, 20, 23-35, 1996.
7. S. Chi, Y-.H. Oh, "Lombard Effect compensation and noise suppression for Noisy Lombard Speech Recognition", Proc. ICASSP, 2013-2016, 1996.
8. S.-Y. Woo, *Robust feature extraction using Lombard effect compensation filter, Master's thesis*, (KAIST, 2003).
9. J. C. Junqua, Y. Anglade, "Acoustic and perceptual studies of Lombard speech: Application to isolated words automatic speech recognition," IEEE Trans., 1990.

## [Profile]

***Sunhee Kim**

Sunhee Kim received the B.A. degree in French Language and Literature from Yonsei University in 1985, the M.S. degree in Linguistics from Université Paris 7 in 1986, and the Ph.D. degree in linguistics from the Ecole Des Hautes Etudes En Sciences Sociales in 1990. From March 1991 she taught French, Linguistics, Phonetics, Phonology and Spoken Language Processing at several universities including Yonsei University. She worked at Lernout & Hauspie Korea as a Senior Researcher from June 2000 to June 2001. She also worked at Speech Information Technology Processing Center of Kwangwoon University as a research professor from March 2002 to February 2004. She joined the Department of Electrical Engineering at KAIST as a research professor in September 2004.

***Soyoung Park**

Soyoung Park received the B.S. degree in Electrical Engineering and Computer Science from KAIST, Korea, in 2004. She is currently taking her Master' courses at the Department of Electrical Engineering and Computer Science at KAIST.

***Chang D. Yoo**

Chang D. Yoo received the B.S. degree in Engineering and Applied Science from California Institute of Technology in 1986, the M.S. degree in Electrical Engineering from Cornell University in 1988 and the Ph.D degree in Electrical Engineering from Massachusetts Institute of Technology in 1996. From January 1997 to March 1999 he worked at Korea Telecom as a Senior Researcher. He joined the Department of Electrical Engineering at KAIST in April 1999.