CEO Clinical and Experimental Otorhinolaryngology Vol. 5, No. 2: 68-73, June 2012 http://dx.doi.org/10.3342/ceo.2012.5.2.68

Original Article

How Does Speaking Clearly Influence Acoustic Measures? A Speech Clarity Study Using Long-term Average Speech Spectra in Korean Language

Heil Noh, MD[1] · Dong-Hee Lee, MD[2]

[1]Department of Otolaryngology-Head and Neck Surgery, St. Vincent's Hospital, The Catholic University of Korea School of Medicine, Suwon; [2]Department of Otolaryngology-Head and Neck Surgery, Uijeongbu St. Mary's Hospital, The Catholic University of Korea School of Medicine, Uijeongbu, Korea

Objectives. To investigate acoustic differences between conversational and clear speech of Korean and to evaluate the influence of the gender on the speech clarity using the long-term average speech spectrum (LTASS).

Methods. Each subject's voice was recorded using a sound level meter connected to GoldWave program. Average long-term root mean square (RMS) of one-third octave bands speech spectrum was calculated from 100 to 10,000 Hz after normalizing to 70 dB overall level using the MATLAB program. Twenty ordinary Korean were compared with 20 Korean announcers with equal numbers of men and women in each group.

Results. Compared with the LTASS of ordinary men, that of ordinary women was lower at low frequencies, but higher at 630, 800, 1,600, 5,000, and 10,000 Hz. Compared with the LTASS of male announcers, that of female announcers was lower at low frequencies. Compared with the LTASS of ordinary men, that of male announcers was significantly lower at 100, 125, 200, and 250 Hz. Compared with the LTASS of ordinary women, that of female announcers was lower at 100, 125, 160, 200, 250, 500, and 10,000 Hz. The LTASS of announcer showed lower levels at 100, 200 Hz and higher at 500, 630, 800, and 1,000 Hz that that of ordinary Koreans.

Conclusion. This study showed that the drop-off of the LTASS in the low frequency region might make the ratings of women and announcers more clearly than those of men and ordinary persons respectively. This drop-off in the low frequency might result in less upward spread of masking and clearer speech. This study reduced an error resulting from a wide variability of clear speech strategies and intelligibility gains, because this study recruited professional speakers. We hope that our results demonstrate the difference in acoustic characteristics of the speech of ordinary Korean persons.

Key Words. *Long-term average speech spectrum, Clarity, Speech, Announcer, Korean*

INTRODUCTION

In everyday communications, the goal of speaking as well as hearing is to communicate someone's message in a manner that is intelligible to others. To speak clearly is very important for the communication among persons who do hear normally as well as those who do not hear normally. It is more important for the communication between person with normal hearing and with hearing impairment. For hearing aids user or cochlear implanted person, some environment can influence the performance of hearing aids or cochlear implants and increase the speaker's variation observed in the physical processes associated with sound production; for example, an important source of variation is the adaptation that speakers make to maintain speech clarity when conversing in noisy environments or over weak mobile phone connections. Therefore, studying the speech clarity can make ev-

• Received October 12, 2011
Revision November 22, 2011
Accepted December 5, 2011

• Corresponding author: **Dong-Hee Lee, MD**
Department of Otolaryngology-Head and Neck Surgery, Uijeongbu St. Mary's Hospital, The Catholic University of Korea School of Medicine, 271 Cheonbo-ro, Uijeongbu 480-717, Korea
Tel: +82-31-820-3564, Fax: +82-31-847-0038
E-mail: leedh0814@catholic.ac.kr

eryday communication easier for many persons (1, 2).

Several acoustic and articulatory factors influencing the speech clarity have been reported in English (3-5). First factor is the speed of the speech; clear speech is slower than conversational speech, which is due to a combination of increased articulation and pause. Second factor is the intensity of the speech; clear speech is produced with greater overall intensity than conversational speech. Third factor is the articulatory features of the articulation organs and vocal tract, which can influence the vowel formant frequency. It was also reported that gender is a remarkable characteristic feature which can affect the speech clarity.

Current understanding about the speech clarity or intelligibility is very limited, especially for Korean. Kwon (6) studied the difference based on gender with respect to acoustic characteristics related with the speech intelligibility to set up the standard data for the future study in various fields in prosthodontics. Although this is the only the study on the speech clarity of Korean language, they just gathered the speech data of both genders and compared the speech intelligibility test scores and several acoustic parameters between male and female subjects. Therefore, from their study, we could get just information about the gender difference related to the speech clarity.

The purpose of this study was to investigate any acoustic difference between conversational and clear speech of Korean. We also compared those data between male and female subjects and tried to find out the influence of the gender on the speech clarity because the gender is widely-known factor of the speech clarity.

MATERIALS AND METHODS

Subjects

Equal numbers of male and female subjects were recruited because speech spectra are known to show sex differences in low- and high-frequency bands (7). Ten men and ten women, aged 24 to 50 years, were enrolled into the study (mean ages±SD: women, 31.9±6.3 years; men, 32.1±7.5 years). All of them lived in the capital area of Korea and spoke standard Korean as their first language. None had any obvious speech defects. To compare the difference between ordinary speakers and professional announcers, ten male and ten female Korean speakers aged between 24 and 48 years were recruited from an announcer school (women, 28.4±6.9 years; men, 27±1.6 years). There were no significant differences in mean age between the four groups found by analysis of variance (ANOVA; P=0.314).

Speech materials

A passage from a novel was chosen on the basis that they were relatively easy to read and did not contain excessive repetition. For most speakers, the material took about 120 seconds to read and provided more than 90 seconds of speech.

Recording equipment

Measurements were obtained at 125 ms intervals using an integration time equivalent to the 'Fast' response of a sound level meter (8). The output of the audiometer microphone was sent to the computer's input. The vocal output at the microphone was monitored using a sound level meter. The frequency responses of the microphone were flat through 10,000 Hz.

Recording procedure and equipment for digital recording stage

Recordings were performed in a soundproofed office (4.2 m wide by 4.8 m long by 2.2 m high) with a noise level of less than 35 dBA. The recording microphone (dynamic univocal microphone; frequency response flat through 10,000 Hz and smooth, peak-free frequency of 50-17,000 Hz) from GenRad 1982 Precision Sound Level Meter (IET Labs Inc., Westbury, NY, USA; weighting flat; range, 30 to 80 dB) was placed on a stand in front of the speaker, at a distance of 20 cm in the same horizontal plane as that of the speaker's mouth and at an azimuth of 0° incidence relative to the axis of the mouth. The speech material was placed in front of the speaker at a distance of 30 cm slightly below the recording microphone to avoid possible reflection from the paper. The speaker was instructed to read aloud at a normal speed and loudness. The vocal output directed toward the microphone was monitored by an assistant observer monitoring a sound level meter. The signals from the recording microphone were transferred to a computer and digitally recorded at a 44.1 kHz sampling frequency and 16 bits resolution using GoldWave ver. 5.2 (GoldWave Inc., St. John's, NL, Canada). While the recorded speech samples were replayed, the boundaries of the sentences were identified acoustically, extracted, and concatenated visually for later analysis of the long-term spectra. Afterward, the recorded traces were digitally high-pass filtered at 100 Hz to remove any background noise that might have come from the equipment.

Analysis stage

This study used acoustic analysis using the long-term average speech spectrum (LTASS) because it is a direct measure of the mean sound pressure level of speech as a function of frequency. It provides a global measure of the acoustic characteristics of continuous discourse in a language (7).

Selection of program to calculate relative root mean square (RMS) level

After obtaining the average RMS spectrum over an integration time of 90 seconds, the RMS pressure in each band was calculated for each time interval. These short interval RMS values were then integrated across the total sampled interval for each speaker. In this process, three different programs were set to determine more suitable logic to calculate relative RMS level. The first program calculated the mean RMS value within each 125 ms inter-

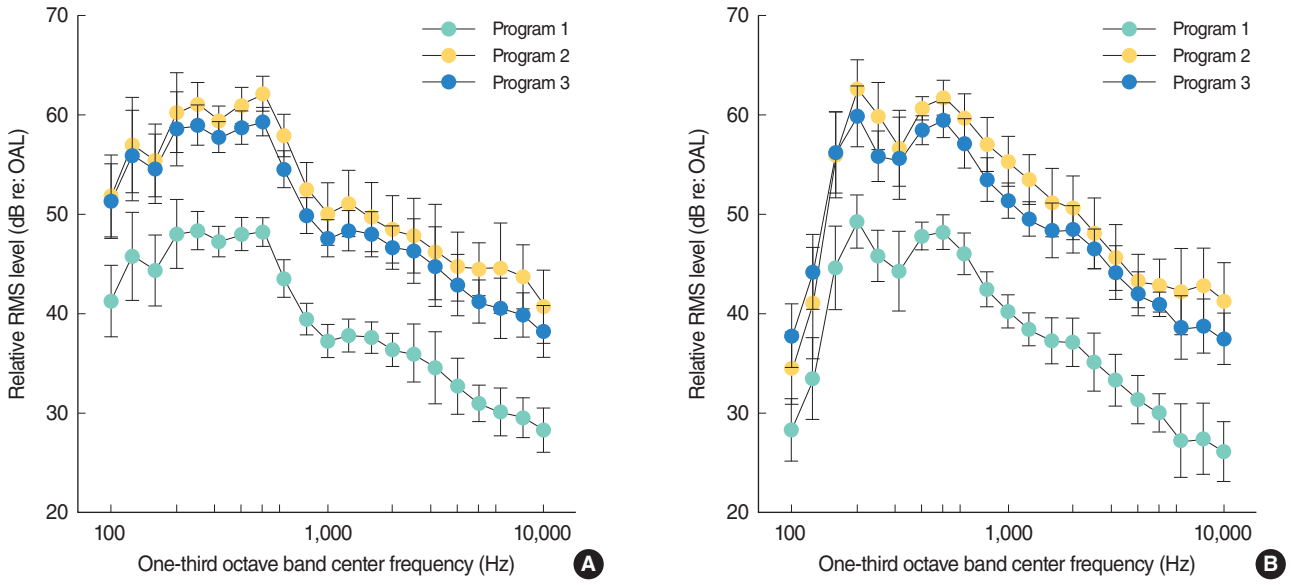

Fig. 1. Error bar graph of relative root mean square (RMS) levels as a function of one-third octave band center frequency. (A) English-speaking men and (B) English-speaking women. In this study, program 3 was used, which chose the maximal RMS value within each 125 ms and then all of the results were averaged through 90 seconds. Circles and bars represent the mean and standard deviation, respectively.

val and all of these were averaged through 90 seconds (program 1). The second chose the maximal RMS value within each 1 second interval and all of these were averaged through 90 seconds (program 2). The third chose the maximal RMS value within each 125 ms and then all of them were averaged through 90 seconds (program 3). To obtain the mean value of each group, the long-interval estimates were averaged across speakers, converted to spectrum levels, and plotted as shown in Fig. 1. The sentences were analyzed in 21 one-third octave bands (over a frequency range of 100-10,000 Hz) using the Infinite Impulse Response (IIR) filter function in the MATLAB program (R2007b, MathWorks, Natick, MA, USA). Program 1 could be influenced by different pauses of discourses and program 2 could overestimate the short-term effects of high-frequency phonemes. Therefore, program 3 was chosen to be the appropriate method for subsequent data collection in this study.

Measurement of LTASS

After calculating the maximal RMS value and normalizing it to an overall level of 70 dB, the long-term average RMS values of one-third octave band speech spectra were calculated for frequencies ranging from 100 to 10,000 Hz. After the signals were transferred from the recording microphone to a computer and digitally recorded, the extracted and concatenated speech samples were digitally high-pass filtered at 100 Hz to remove any background noise that might have arisen from the equipment. This part was done by using GoldWave. Before normalizing the individual voice to 70 dB overall level, the average RMS envelope level was calculated during 90 seconds of speech. To calculate the RMS value within 125 ms-moving time window, 90 seconds of speech went through 125 ms-moving rectangular time

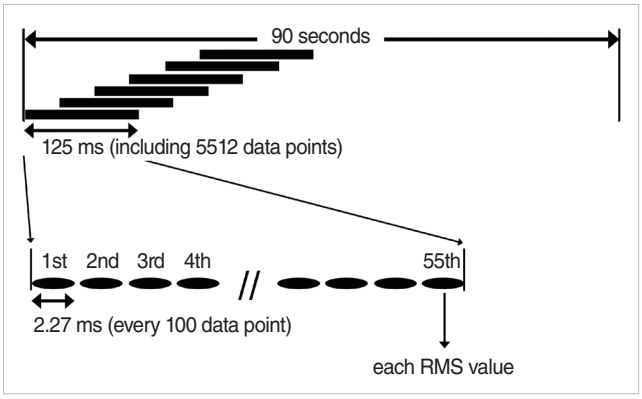

Fig. 2. Diagram showing how to calculate the maximal root mean square (RMS) value and how to normalize it for different voice level in this study.

window and each 125 ms included 5,512 data points (125 ms× 44.1 kHz of sampling frequency). Byrne et al. (9) sampled the RMS envelope at the interval of 31.25 ms, but we chose the interval of 2.27 ms (every 100 data points) to obtain better resolution. Therefore, we could obtain 55 RMS values (5,512÷100) within 125 ms-moving time window and chose the maximal one from 55 RMS values. Seven hundred twenty maximal RMS values of each 125 ms-moving time window were averaged for 90 seconds of speech and averaged value was converted to dB as the overall level of 90 seconds speech sample (Fig. 2). The ratio between each overall level and 70 dB was used to normalize the individual voice data to 70 dB overall level. If the louder voice data were 80 dB OAL, all the data points were divided by $10^{(80/20)}$ and multiplied by $10^{(70/20)}$, if the softer voice data were 60 dB OAL, the softer voice data were divided by $10^{(60/20)}$ and multi-

plied by $10^{(70/20)}$.

The sentences were analyzed in 21 one-third octave bands (over a frequency range of 100-10,000 Hz) using the IIR filter function in the MATLAB program. The group data were represented in mean±SD of these data for each 21 one-third octave bands. Each average long-term RMS value of one-third octave band was obtained by summing and averaging components over one-third octave intervals with center frequencies ranging from 100 Hz to 10,000 Hz. The digital filter was designed to match the ANSI standard. As we know, ANSI S1.11-1986 uses Butterworth filters (maximally flat magnitude filter type) to define the order and attenuation of the octave filters. Peak-to-valley ripple in the pass band is less than 0.5 dB with attenuation within ± 0.5 dB. Attenuation of frequencies outside 5 times and 1/5 of the band center frequency is better than 75 dB. So according to these regulations, we designed MATLAB function like this: [b a]=butter (3, [f1 f2].*2/44100, 'bandpass'). Our digital analysis protocol maximally simulated the analogue methods of Byrne et al. (9), who used a sound level meter detecting the RMS envelope of the input signal.

Statistical analysis

The LTASS measures were analyzed separately for men and women because of the differences in their fundamental frequencies and formants. The nonparametric Mann-Whitney *U* test was used to compare the values between pairs of groups and $P<0.05$ was considered statistically significant. All the statistical analyses were performed using IBM SPSS ver. 19 (IBM Co., New York, NY, USA).

Ethical considerations

The study design and experimental protocols were approved by the Institutional Review Board of St. Vincent's Hospital of the Catholic University of Korea.

RESULTS

Comparison between men and women

The LTASS of ordinary women was lower than that of ordinary men at lower frequencies and the difference was statistically significant at 100, 125, 250, and 315 Hz ($P<0.001$, $P<0.001$, $P<0.001$, and $P=0.005$, respectively). But the LTASS of ordinary men was significantly lower than that of ordinary women at 630, 800, 1,600, 5,000, and 10,000 Hz ($P=0.011$, $P<0.001$, $P=0.049$, $P=0.005$, and $P=0.001$, respectively). At other frequencies, the LTASS of ordinary men and women was similar.

The LTASS of female announcers was lower than that of male announcers at lower frequencies and the difference was statistically significant at 100, 125, 160, 315, and 1,250 Hz ($P<0.001$, $P<0.001$, $P<0.001$, $P<0.001$, and $P=0.005$, respectively). In announcer groups, the LTASS of male and female subjects was similar at other frequencies and there was no frequency at which the LTASS of men was lower than that of women.

Comparison between announcers and ordinary persons

The LTASS of male announcers was generally higher than that of ordinary men and the difference was statistically significant at 160, 315, 630, 800, and 1,000 Hz. The exceptions are 100, 125, 200, and 250 Hz, where the LTASS values of male announcers were found to be significantly lower than those of ordinary men (Fig. 3A).

There were no significant differences in the LTASS along all frequencies except at 100, 125, 160, 200, 250, 500, and 10,000 Hz between ordinary women and announcers (Fig. 3B).

When the combined data of men and women were compared, the LTASS of announcer showed significantly lower levels at 100, 200 Hz and higher at 500, 630, 800, and 1,000 Hz. There was no significant difference in the LTASS above 1 kHz, even though the examiners perceived that the announcers' speech was more clearly pronounced (Fig. 3C).

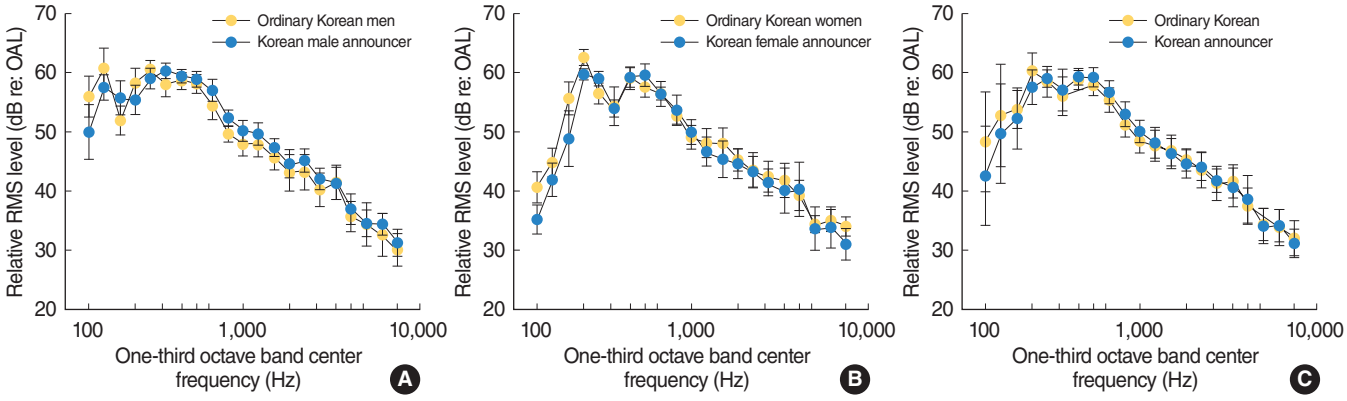

Fig. 3. Error bar graph of relative root mean square (RMS) levels as a function of one-third octave band center frequency. Novel reading by ordinary Korean speakers and Korean announcers. (A) Men, (B) women, and (C) combined data of men and women. Circles and bars represent the mean and standard deviation, respectively.

DISCUSSION

According to source-filter theory of speech, speech production can be divided into two independent parts; the source of sound is the larynx and the filter that modifies the sound is the vocal tract (10). There are many differences between consonants and vowels. Consonants have rapid changes in articulators; they are produced by making constrictions in the vocal tract and they need coordination of all three sources (frication, aspiration, and voicing). They decide the intensity (loudness cues), frequency (pitch cues), and duration (length of sound cues) of the speech. Vowels have slow changes in articulators; they are produced by a relatively open vocal tract and only the voicing source is used. Contrast to consonants, vowel resonances tend to have a bandwidth of low frequency. Generally, center-frequency of F1 and F2 of vowels are known to be approximately as follows: for /i/, F1=200-300 Hz, F2=2,000-3,000 Hz; for /e/, F1=400-700 Hz, F2=1,700-2,200 Hz; for /a/, F1=500-1,000 Hz, F2=1,000-1,700 Hz; for /o/, F1=400-700 Hz, F2=700-1,000 Hz; for /u/, F1=200-300 Hz, F2=500-900 Hz. The acoustic and articulatory features of vowels have been known to be a key element for speech clarity. First, vowels tend to have longer durations and an expanded formant space (11). Although consonant context effect on vowel formants is similar for both clear and conversational speech, clearly produced vowels are much longer in duration. Second, the size of vowel formant space increases with the clarity of the speech.

There has been the one and only report which studied the speech clarity or intelligibility of Korean (6). Therefore, this study is the second one but there are many differences between two studies.

The gender is among the best known factors influencing speech clarity or intelligibility (4-6). Kwon (6) compared the conversational speech data of men with that of women in terms of the speech intelligibility and set up the data for the future study in various field in prosthodontics. They compared several acoustic parameters of the speech according to the gender but they collected just conversational speech of ordinary persons to evaluate the speech intelligibility. Their methods had a limitation in evaluating the factors influencing the speech intelligibility as well as the gender difference in acoustic features because they did not compare the conversational speech with clear one. However, we gathered the clear speech of professional speakers who were trained to be an announcer and compared the data of them with those of ordinary persons. This methodology is important because of a wide variability of clear speech strategies and intelligibility gains that are presented to speakers by the researcher in the recording situation. In most studies, including ones by Kwon (6), both conversational and clear speech is the speech elicited by specific instructions given to speakers rather than to the spontaneous speech occurring in a more natural setting like professional speakers. The instructions most typically involve asking speakers to read the same set of materials twice or more in the ambiguous manner of the following: "read the materials as if you were talking to someone with a hearing impairment" or "speak clearly and precisely", etc. In those cases, talkers tend to interpret the task in a way that they saw appropriate but unproved by themselves. This resulted in a wide variability of clear speech strategies and intelligibility gains (4). Because the clear speech of professional speakers who were trained to be an announcer was analyzed in this study, we could reduce this error.

In the research methodology about the speech clarity, it should be considered which category of the speech measurement is analyzed: global or segmental measurements. Kwon (6) measured the segmental parameters including first formant, second formant and fundamental frequency of three vowels /a/, /i/, and /u/, and evaluated the vowel working space of three vowels /a/, /i/, and /u/. However, we analyzed the representative global parameter of the speech, LTASS. The LTASS is a direct measure of the mean sound pressure level of speech as a function of frequency. Because it can provide a global measure of the acoustic characteristics of continuous discourse in a language, it has been widely used in the prescription and evaluation of hearing aid fittings (7, 12).

Until now, many acoustic studies reveal that the factors of the speech clarity typically involve a wide range of acoustic/articulatory adjustments, including a decrease in speaking rate, wider dynamic pitch range, greater sound pressure levels, more salient stop releases, greater RMS intensity of the non-silent portions of obstruent consonants, increased energy in 1-3 kHz range of long-term spectra, higher-voice intensity, vowel space expansion, and vowel lengthening (3, 4). Slow speaking rate means longer segments as well as longer and more frequent pauses of the speech, not merely the speed of speaking. Although numerous acoustic-phonetic features of clear speech have been identified, it is not well-understood yet how each of these modifications affects the speech clarity or intelligibility. For example, female talkers tend to produce more intelligible and clear speech compared to male talkers but we do not know how above numerous acoustic-phonetic factors result in making the female speech clearer.

This study showed that the LTASS of women was significantly lower than that of men at low frequencies and the LTASS of announcer showed significantly lower levels at low frequencies than that of ordinary persons. That means that the drop-off of the LTASS in the low-frequency region might make the ratings of women and announcers more clear than those of men and ordinary persons respectively. This drop-off in the low frequency might result in less upward spread of masking and clearer speech. It has been known that reduction in low-frequency gain can improve the speech recognition in noise (13). Two mechanisms about this phenomenon was assumed. One is that reduction of gain in low-frequency region can improve the signal-to-noise ratio (SNR) because the background noise contains predominantly the energy in this region. The other is reduction of low-frequen-

cy gain can alleviate the upward spread of masking. For most environmental background noises are broadband, the latter assumption is more preferable. In this study, all of recordings were performed in a soundproofed office with a noise level of less than 35 dBA and this effect might primarily attribute to clearer speech of announcers. However, the readers should keep in mind that this effect on clear speech may be less in most everyday life, in which most noise contains low-pass noise as well as broadband noise. In this circumstance, other factors like a decrease in speaking rate, greater sound pressure levels, higher voice intensity, or more salient stop releases may be dominant.

Generally, female speakers exhibit significantly greater acoustic specification than do male speakers. This finding may be due to vocal tract size differences between men and women (14). The gender difference in vocal tract size can influence the vowel formant frequency (15). Because women have smaller vocal tracts than men, an equivalent degree of articulatory specification may produce relatively greater acoustic specification in women than in men. This is also why the vowels of female speakers are more intelligible during clear speech in noise compared with the vowels of male speakers (16, 17).

According to Leino's study (18), the good and ordinary voices differed from the poor ones in their relatively higher sound level in the frequency range of 1-3 kHz and a prominent peak at 3-4 kHz. He also commented that this finding was not dependent on articulation and thus most likely was not language dependent either, because this peak at 3.5 kHz can also be seen in the LTASS of speakers of French, Danish, German and English. However, because the origin of Korean language is different from those of Latin languages like French or Germanic languages like Danish, German, and English, we should keep in mind that there may be any phonological difference between Korean and other languages.

In conclusion, this study showed that the drop-off of the LTASS in the low frequency region might make the ratings of women and announcers more clearly than those of men and ordinary persons, respectively. This drop-off in the low frequency might result in less upward spread of masking and clearer speech. We hope that our results demonstrated the difference in acoustic characteristics of the speech of ordinary Korean people and professional speaker and help the readers better understand the acoustic characteristics of the clearly speaking speech in Korean.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

ACKNOWLEDGMENTS

The authors thank Mu Gil Heo (Jeil Hearing Aid Center, Seoul, Korea), who helped the measurement of LTASS. The authors thank the Navi Speech and Image School's announcers, who participated in the recording of the speech materials.

REFERENCES

1. Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing I: intelligibility differences between clear and conversational speech. J Speech Hear Res. 1985 Mar;28(1):96-103.
2. Payton KL, Uchanski RM, Braida LD. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. J Acoust Soc Am. 1994 Mar;95(3):1581-92.
3. Tasko SM, Greilick K. Acoustic and articulatory features of diphthong production: a speech clarity study. J Speech Lang Hear Res. 2010 Feb;53(1):84-99.
4. Smiljanic R, Bradlow AR. Speaking and hearing clearly: talker and listener factors in speaking style changes. Lang Linguist Compass. 2009 Jan;3(1):236-64.
5. Bradlow AR, Torretta GM, Pisoni DB. Intelligibility of normal speech. I: global and fine-grained acoustic-phonetic talker characteristics. Speech Commun. 1996 Dec;20(3):255-72.
6. Kwon HB. Gender difference in speech intelligibility using speech intelligibility tests and acoustic analyses. J Adv Prosthodont. 2010 Sep;2(3):71-6.
7. Cox RM, Moore JN. Composite speech spectrum for hearing and gain prescriptions. J Speech Hear Res. 1988 Mar;31(1):102-7.
8. Olsen WO, Hawkins DB, Van Tasell DJ. Representations of the long-term spectra of speech. Ear Hear. 1987 Oct;8(5 Suppl):100S-108S.
9. Byrne D, Dillon H, Tran K, Arlinger S, Wilbraham K, Cox R, et al. An international comparison of long-term average speech spectra. J Acoust Soc Am. 1994 Oct;96(4):2108-20.
10. Titze IR. Nonlinear source-filter coupling in phonation: theory. J Acoust Soc Am. 2008 May;123(5):2733-49.
11. Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. J Speech Hear Res. 1986 Dec;29(4):434-46.
12. Byrne D, Dillon H. The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid. Ear Hear. 1986 Aug;7(4):257-65.
13. Cook JA, Bacon SP, Sammeth CA. Effect of low-frequency gain reduction on speech recognition and its relation to upward spread of masking. J Speech Lang Hear Res. 1997 Apr;40(2):410-22.
14. Mefferd AS, Green JR. Articulatory-to-acoustic relations in response to speaking rate and loudness manipulations. J Speech Lang Hear Res. 2010 Oct;53(5):1206-19.
15. Liu H, Ng ML. Formant characteristics of vowels produced by Mandarin esophageal speakers. J Voice. 2009 Mar;23(2):255-60.
16. Bradlow AR, Bent T. The clear speech effect for non-native listeners. J Acoust Soc Am. 2002 Jul;112(1):272-84.
17. Ferguson SH. Talker differences in clear and conversational speech: vowel intelligibility for normal-hearing listeners. J Acoust Soc Am. 2004 Oct;116(4 Pt 1):2365-73.
18. Leino T. Long-term average spectrum in screening of voice quality in speech: untrained male university students. J Voice. 2009 Nov; 23(6):671-6.