Increasing Speech Intelligibility by Mimicking Professional Announcers' Voices and Its Physical Correlates

Dung Kim Tran*, Masato Akagi*, and Masashi Unoki*
* School of Information Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
E-mail: {kimdungtran,akagi,unoki}@jaist.ac.jp

***Abstract*—Previous studies found that speech uttered by professional announcers is more intelligible than that by non-experts in noisy environments. On the basis of this finding, we developed a voice-conversion (VC) system to mimic professional announcers' voices by modifying the speaker embedding of non-expert speech. The results from our experiments to evaluate this system indicated that intelligibility increased significantly with this system. In this paper, to discuss what physical features correlate to the intelligibility, the following two issues are investigated by analyzing this system: (1) whether speech intelligibility can be changed gradually even by shifting one PCA (principal component analysis) component of the speaker embedding of the above VC system and (2) what physical features are changed when the PCA component is shifted, we retrained the VC system with a larger amount of training data. Comparing the speech intelligibility and candidate features that were changed with the shift of one axis of PCA, we found that spectral tilt, spectral plateau, and cepstral peak prominence are strongly correlated with intelligibility.**

Keywords: spectral tilt, spectral plateau, cepstral peak prominence, PCA, STOI, voice conversion

I. Introduction

Speech intelligibility is important in providing general information such as the arrival/departure time of trains and flights; driving instructions, nearby gas stations, and parking lots. It is especially important in announcing emergency situations such as fires, earthquakes, and landslides; as well as guiding people to safety. Making speech more intelligible is specifically challenging in noisy and reverberant environments such as airport/train stations, kitchens, and shower rooms. It is even more difficult when listeners are elderly with or without hearing impairments.

The easiest and most natural response to the challenges is turning up the volume of playback devices; however, increasing the volume is not equivalent to increasing the intelligibility of speech. A person can involuntarily increase the intelligibility of his/her speech in noisy places, known as the Lombard effect [1]. But increasing the volume of playback devices does not have such effect. Various approaches have been proposed to increase speech intelligibility in noisy environments without increasing the total power of speech. These include modification of spectral properties [2]–[4], dynamic range compression [5]–[7], modification of speech modulation spectrum [8], [9], and time-scale modification [10], [11].

Several studies have found that voice-related professions, e.g., professional announcers, voice actors, and singers, can produce speech that is clearer and easier to hear than non-professional people [12], [13]. Other studies have shown that the speech by professional announcers can maintain intelligibility better than that by non-experts in very noisy environments [14]. Inspired by this phenomenon, a previous study [15] developed an end-to-end deep neural network (DNN) based voice conversion (VC) system that mimicked the speaking style of professional announcers to increase speech intelligibility of non-expert people.

As reported in the previous study [15], by analyzing the principal component analysis (PCA) of speaker embedding in the proposed VC system, the first PCA component was found to be related to gender difference and the second PCA component of speaker embedding captured the difference between non-expert speakers and professional announcers. This is advantageous as a non-expert voice can be converted to have the voice style of a professional announcer only by changing the second PCA component of a speaker embedding. Since professional announcers' voices are more intelligible even in a noisy environment, it is expected that the second PCA component can be used to increase the intelligibility of non-expert voices. To clarify this point, the authors [15] proposed replacing the second PCA component of a non-expert speaker embeddings with the average value calculated from the second principal component of all male professional announcers. The obtained speaker embedding was then used to synthesize converted stimuli.

Experimental results [15] from objective measurements and subjective evaluation of the converted stimuli confirmed that adapting to an announcer's voice can increase the intelligibility of a non-expert speaker's voice. By modifying the second PCA component of speaker embedding, we can manually control the announcer speaking style, hence increase the intelligibility of speech in a noisy environment without completely changing speaker individuality. Statistical analysis also showed that modifying the second PCA component yields the highest performance. These results suggest that the second PCA com-

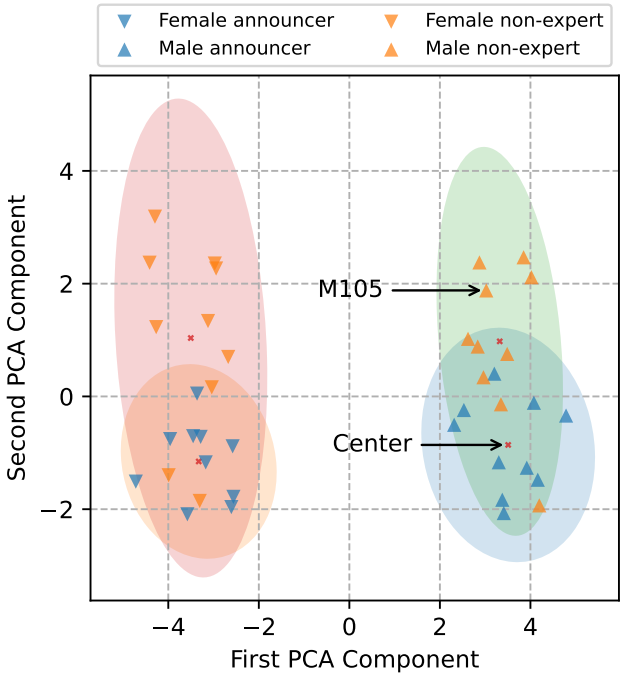

Fig. 1. Speaker embeddings in PCA domain

ponent is related to the intelligibility of speech.

However, the DNN-based VC system is a black box; it is unknown what happens that improves the intelligibility of speech when the second PCA component is swapped. It is also unclear whether speech intelligibility can be changed gradually by shifting only the second PCA component. This paper analyzes speaker embeddings in the system [15] that is thought to be a black box in a DNN-based VC for improving speech intelligibility and discusses physical cues for improving intelligibility. More specifically, we retrain the above VC system with a larger amount of training data and use it to control the second PCA component. We then study the subsequent effects on physical properties and intelligibility of speech and present two discussion points:

- Whether speech intelligibility can be changed gradually even by shifting one PCA component of the speaker embedding.
- What physical features are changed when the PCA component is shifted.

Possible candidates of physical correlates are spectral tilt and plateau and cepstral peak prominence (CPP).

II. Voice-conversion system

A. Training voice-conversion system

StarGANv2-VC [16] is one of the most effective VC systems. On the basis of this system, the authors [15] proposed a method for converting the voice of a non-expert speaker to that of a professional announcer to increase speech intelligibility. Their goal was to increase the effectiveness of StarGANv2-VC by adding automatic speech recognition to the discriminator module. They also used a Parallel WaveGAN [17] vocoder to generate converted speech utterances from the corresponding converted mel-spectrograms.

We retrained the above VC system with a larger amount of training data to discuss the properties of the PCA second component of the speaker embedding related to expert versus non-expert speakers. More specifically, the training data consist of utterances from 20 professional announcers from the ATR dataset A-set and 20 non-expert speakers from the ATR dataset C-set [18]. All the utterances are pre-processed by resampling to 24 kHz, removing leading and trailing silence; and they are combined to 5-second chunks. There are total 29,956 utterances, in which 500 utterances are used for validation. The 80-band log-mel spectrogram with band-limited frequency range (0 to 8 kHz) is extracted using short-time Fourier transform. The window length and frame shift are set to 1024 and 256, respectively.

We follow the training strategy described in the above studies [15], [16] with the same objective functions and hyper-parameters. We then used the style encoder of our retrained system to generate the PCA of the speaker embeddings for non-experts and professional announcers to verify that the second PCA component is related to expert and non-expert speakers.

Figure 1 shows the visualization of the first two PCA components of the speaker embedding. It can be seen that the first PCA component is related to gender with female and male speakers distributed on the left and right. The second PCA component is related to expert and non-expert speakers. The clusters of female/male professional announcers are compact and those of non-expert speakers are wide on the second PCA component. This phenomenon suggests that the professional announcers can control the configurations of their voice organs and high intelligible speech is the result of such effort, resulting that the variance of their voices are smaller than those of non-expert speakers.

B. Shifting second PCA component

We referred to the second PCA component of a non-expert speaker (M105 in the ATR dataset C-set) as the source and average value of the second PCA component of ten professional male announcers in the ATR dataset A-set as the Center. The PCA components of M105 and the Center are visualized in Fig. 1. Let d be the distance between the PCA values (the second PCA component) of the source and Center. We calculated three checkpoints, i.e., C1 $= 0.25d$, C2 $= 0.5d$ and C3 $= 0.75d$ from the source, respectively. For example, assuming that the second PCA components of the source and Center were vectors: $\mathbf{a}$ and $\mathbf{b}$; the checkpoints were calculated as: C1 $= 0.25d = (1-0.25)\times\mathbf{a}+0.25\times\mathbf{b}$, C2 $= 0.5d = (1-0.5)\times\mathbf{a}+0.5\times\mathbf{b}$, and C3 $= 0.75d = (1-0.75)\times\mathbf{a}+0.75\times\mathbf{b}$.

With our retrained voice-conversion system, we first calculated the PCA components for a non-expert speaker and all ten male professional announcers. We then calculated the Center by taking the average of the second PCA components corresponding to the professional announcers and calculated

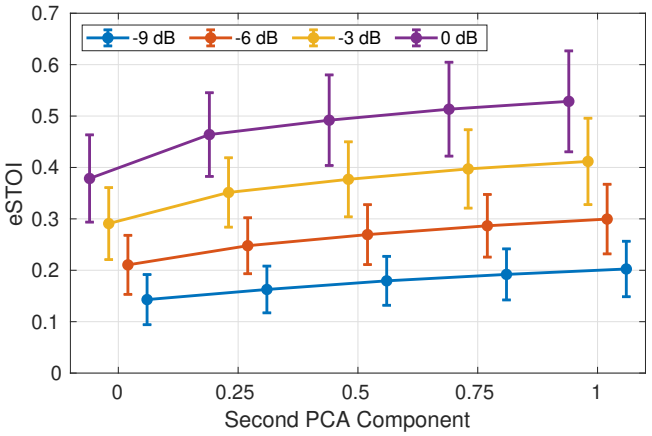

Fig. 2. eSTOI scores at four levels of noise (-9, -6, -3, 0 dB). On the horizontal axis, 0 denotes the source and 1 is the Center. 0.25, 0.5 and 0.75 are the checkpoints, C1, C2, and C3, respectively.

the distance between the second PCA components of the non-expert speaker and the Center as d. We then used d to calculate the PCA values for checkpoints C1, C2, and C3 accordingly. Finally, in turns, we replaced the second PCA component of the non-expert speaker embedding with the PCA values of C1, C2, C3, and Center and used our retrained voice-conversion system to produce the converted speech stimuli.

III. Second PCA component and intelligibility in noise

The first discussion point is investigated in this section. We conducted an experiment to gain insights about the correlation between the second PCA component and speech intelligibility. The consistency of the correlation was also examined in various levels of noise.

We selected 520 Japanese words, each containing 1 to 4 morae, from the ATR dataset A-set and ATR dataset C-set as the clean stimuli for conversion-target and source speakers. All speech waveforms were pre-processed to a 16-kHz sampling rate with a single channel. We followed the instructions in [15] and used 16 kHz here instead of 24 kHz as described in Section II. There were 6 types of speech stimuli in the experiments: S_N is the natural speech of a non-professional speaker, collected from speaker M105 in the C-set; S_A is the natural speech of professional announcers, collected from 20 male and female speakers in the A-set; and S_1, S_2, S_3, and S_C are converted speech produced by shifting the second PCA components of speaker embedding of the source (M105) to checkpoint C1, C2, C3, and Center, respectively.

To create the noisy stimuli, we masked the clean stimuli with pink noise at four different signal-to-noise-ratio (SNR) levels: -9, -6, -3, and 0 dB. We calculated the root-mean-square of the speech signal only in the speech-presence region and scaled the noise signal to match the desired SNR level. The speech-presence region is derived from the text labels of the ATR dataset. To avoid the effect of different onset and offset timing between speech stimuli, the duration of each stimulus was adjusted to contain the same 200 ms of leading noise and 200 ms of trailing noise. Speech stimuli were gated with two raised cosine onset and offset windows of 40 ms to avoid overshoot distortion.

To objectively measure the intelligibility of speech in noise and determine whether speech intelligibility is increased gradually by shifting values along the second PCA component, we calculate the extended Short Term Objective Intelligibility (eSTOI) of the speech stimuli at 4 SNR levels: -9, -6, -3 and 0 dB at 5 checkpoints (Source, C1, C2, C3, and Center) using pySTOI python package [19]. The clean speech by speaker M105 were used as the reference signals at checkpoint 0 and the clean-converted speech were used as the reference signals at other checkpoints for eSTOI calculation.

Figure 2 shows the experimental results. The vertical lines at each checkpoint correspond to the standard deviations of eSTOI scores. It appears that intelligibility was increased gradually when the second PCA component was shifted from source to checkpoints C1, C2, C3, and Center. The eSTOI scores were also consistent with various levels of noise. The results provide an important insight in that the intelligibility could be tuned adaptively to various levels of noise on the basis of the voice of an arbitrary person by shifting the second PCA component using our retrained system.

IV. Intelligibility and its physical correlates

A. Spectral tilt, spectral plateau, and intelligibility

Studies on laryngeal anomalies have found that breathiness is a prominent feature used for determining various pathological conditions [20]. Breathiness is a phenomenon caused by the incomplete closure of vocal fold. As a result, breathy voices are more difficult to hear (less intelligible) in noisy environments. Compared with normally phonated signals, breathy glottal signals have steeper downward spectral slopes, less energy in the high-frequency region, and are less periodic [21]. Several acoustical studies have reported correlations between spectral tilt (downward spectral slope), spectral plateau (energy in the high-frequency region); and breathiness ratings [22]. Based on these knowledge, we hypothesize that a speech signal that has a flatter spectral tilt and larger plateau energy could have higher intelligibility and vice versa.

Figure 3 (a) shows an example of the spectral tilt we calculated for this study. It is the frequency spectrum of the vowel /a/ with the spectral tilt presented as the red line. We used the label of the ATR dataset to extract a segment of vowel /a/. We then calculated the Fourier spectrum of the segment. Finally, we calculated the regression line of the spectrum between 300 and 4,000 Hz. We used the slope of the regression line as the spectral tilt. Figure 3 (b) shows an example of the spectral plateau. On the same spectrum shown in Fig. 3 (a), the spectral plateau is presented as the red line between 2,000 and 4,000 Hz. We subtracted the entire spectrum from its peak magnitude. We then calculated the average magnitude of the spectrum in the 2,000-4,000-Hz band. We used this average magnitude to represent the spectral plateau and compared it with the plateaux of other spectra.

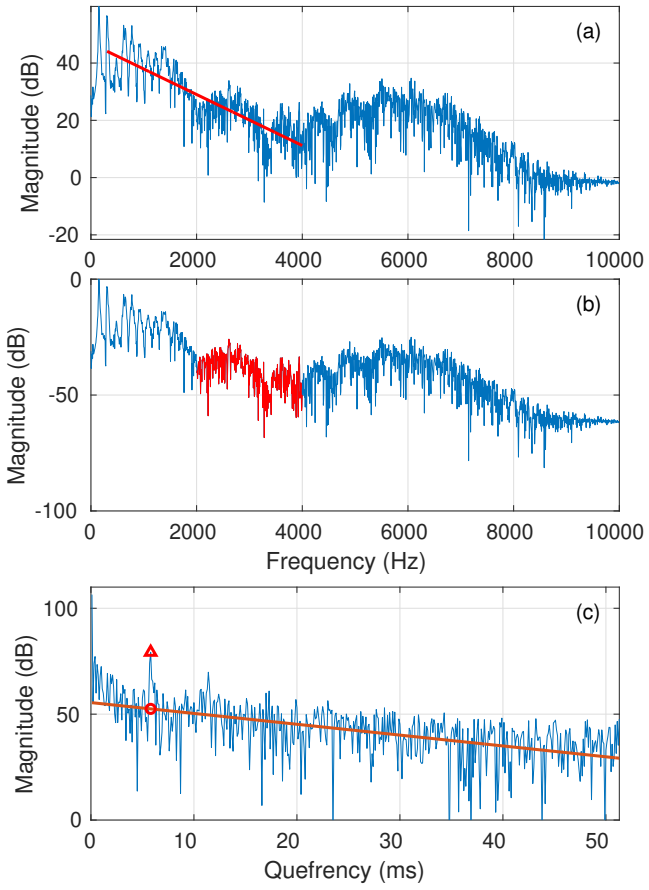

Fig. 3. Frequency spectrum and quefrency cepstrum of vowel /a/

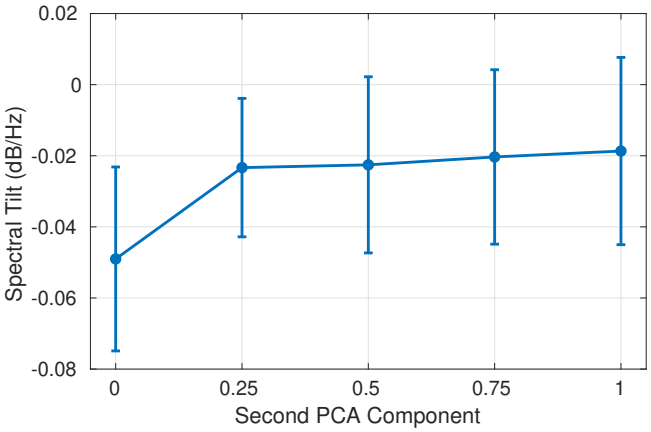

Fig. 4. Average spectral tilt and standard deviation of vowels /a/, /e/, /i/, /o/, /u/. On the horizontal axis, 0 denotes the source and 1 is the Center. 0.25, 0.5 and 0.75 are the checkpoints, C1, C2, and C3, respectively.

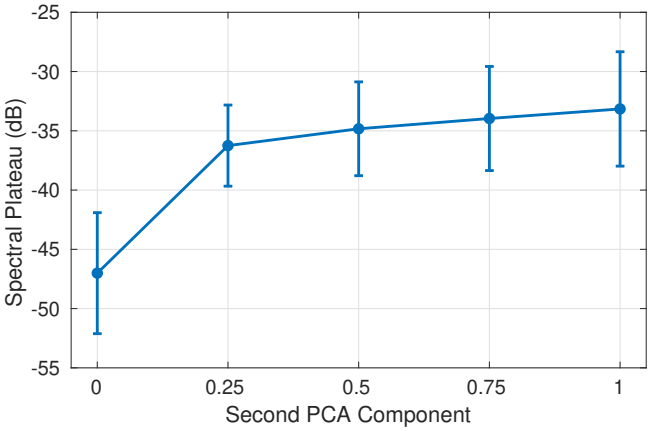

Fig. 5. Average spectral plateau and standard deviation of vowels /a/, /e/, /i/, /o/, /u/. On the horizontal axis, 0 denotes the source and 1 is the Center. 0.25, 0.5 and 0.75 are the checkpoints, C1, C2, and C3, respectively.

B. Periodicity, CPP, and intelligibility

A breathy voice is less periodic compared with a normal one and the quefrency domain is effective for analyzing signal periodicity. A crude explanation of the quefrency domain is that we treat the spectrum of a speech waveform as a signal and apply Fourier transform to it; the result is the cepstrum of the speech waveform in the quefrency domain. It can be observed that a highly periodic speech spectrum should result in clear harmonics in the quefrency domain; therefore, clearer cepstral peaks.

Several analyses have found a strong relation between the periodicity of a signal and its CPP [20], [23]. A CPP is the distance between the most prominent peak and the level of the cepstral background noise on the regression line immediately below the peak. Experimental results [20]–[23] indicated that the CPP of a breathy voice is significantly lower than that of a normal voice. In other words, a higher CPP level is an indicator of higher speech intelligibility. Figure 3 (c) shows an example of the CPP in the quefrency domain. It is the cepstrum of the vowel /a/ with its regression line and CPP shown as the red line and red triangle, respectively. We use the regression line of the cepstrum as the referenced background noise. The CPP is calculated as the difference in magnitude of the highest peak (the red triangle) and the red circle directly under it on the regression line.

C. Second PCA component, physical correlates, and intelligibility

Thus far, we have learned from the literature that a shallower spectral tilt, larger amount of energy of the spectral plateau, and higher CPP are associated with higher speech intelligibility. The next step is examining the correlation between the second PCA component and acoustical properties. We shifted the second PCA component, as described in Section II-B, and used our retrained voice conversion system to generate five types of stimuli. We then used the ATR dataset label to extract the vowel /a/, /e/, /i/, /o/, /u/ segments of five types of stimuli (S_N, S_1, S_2, S_3, and S_C). Finally, we used these segments to calculate spectral tilt, spectral plateau, and CPP.

The results of this experiment are shown in Figs. 4, 5, and 6. The vertical lines at each checkpoint correspond to the standard deviations of the averaged spectral tilts, spectral plateaux, and

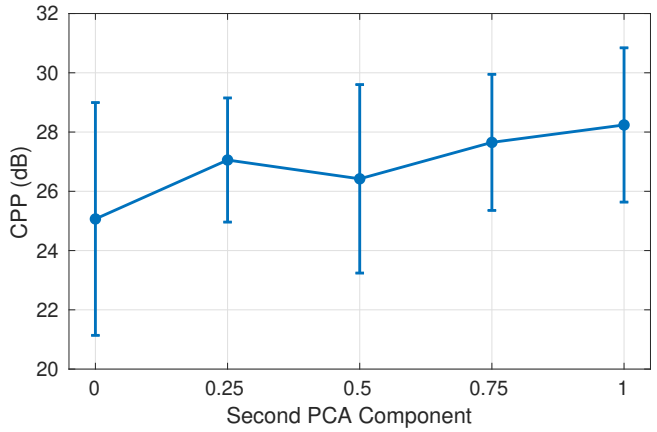

Fig. 6. Average cepstral peak prominence and standard deviation of /a/, /e/, /i/, /o/, /u/. On the horizontal axis, 0 denotes the source and 1 is the Center. 0.25, 0.5 and 0.75 are the checkpoints, C1, C2, and C3, respectively.

CPPs. It is known that a breathy voice (low intelligibility) has a steeper spectral slope (steeper tilt) and less energy in the high-frequency region (small or no plateau). It can be seen in Fig. 4 that the averaged slopes of the spectral tilts are reduced gradually from about -0.05 dB/Hz at checkpoint 0 to above -0.02 dB/Hz at the Center. Similarly, it appears in Fig. 5 that the average energy in the high-frequency regions (spectral plateaux) are rised from about -47 dB at checkpoint 0 to about -33 dB at the Center. Observing the experimental results visualized in Figs. 4, 5, and 2, there is a relation between the spectral tilt, spectral plateau, and intelligibility, i.e., the improvement of these physical properties is consistent with the improvement of speech intelligibility. More specifically, the non-expert speaker has the steepest spectral tilt and smallest spectral plateau, suggesting that the non-expert speaker has the lowest speech intelligibility. In contrast, the Center appears to have the shallowest spectral tilt and largest spectral plateau; therefore, it has the highest intelligibility. Between the non-expert speaker and Center, checkpoints C1, C2, and C3 have a small fluctuation; however, the overall trend is that the spectral tilts become shallower and spectral plateaux larger, indicating that the intelligibility becomes higher when the second PCA component is shifted from the non-expert speaker to checkpoints C1, C2, C3, and finally the Center.

It has been reported that the spectrum of a breathy voice (low intelligibility) is less periodic, and a less periodic spectrum results in a less prominent cepstral peak. It can be seen in Fig. 6 that the averaged cepstral peaks have become more prominent, rising from about 25 dB at checkpoint 0 to above 28 dB at the Center. Observing the experimental results visualized in Figs. 6 and 2, there is a relation between the CPP and intelligibility, i.e., the improvement of the CPP is consistent with the improvement of speech intelligibility. More specifically, the non-expert speaker has the lowest CPP; in other words, the speaker has the least speech intelligibility. The CPP values of other checkpoints are all larger than that of the non-expert. Although there is a small fluctuation, it appears that the CPP values becomes larger, indicating that the intelligibility becomes higher when the second PCA component is shifted from the non-expert speaker toward the Center.

The experimental results visualized in Fig. 2 show that speech intelligibility is increased gradually when the second PCA component is shifted gradually. The shift also produces enhancements of spectral tilt, spectral plateau, and CPP as shown in Figs. 4, 5, and 6, respectively. These results provide evidence that there is a correlation between the improvement of intelligibility and enhancement of physical properties of speech.

V. Conclusion

We presented two points of discussion for finding the correlations between the second PCA component, physical properties, and intelligibility of speech in order to learn why speech intelligibility can be increased by mimicking professional anouncers' voices using the end-to-end DNN-based VC system [15]. The first point is whether speech intelligibility can be changed gradually even by shifting one second PCA component of the speaker embedding and the second point is what physical features are changed when the second PCA component is shifted. To do so, we retrained the VC system [15] with a larger amount of training data. We then used our retrained system to shift the second PCA component of a non-expert speaker to those of checkpoints C1, C2, C3, and the Center. We used the eSTOI to evaluate the intelligibility of speech stimuli with four levels of noise (-9, -6, -3, and 0 dB) at five checkpoints. Experimental results confirm for the first point of discussion that speech intelligibility can be increased gradually by shifting the second principal component toward those of professional announcers.

We investigated why it is possible to increase speech intelligibility by using the second PCA component of professional announcers. We looked into the spectral tilt and spectral plateau in the frequency domain, as well as the CPP in the quefrency domain. The results confirm for the second point of discussion that when the second PCA component is shifted from a non-expert speaker through checkpoints C1, C2, C3, and Center, all the acoustical properties related to intelligibility appear to be enhanced proportionally.

VI. Acknowledgements

This work was supported by the SCOPE Program of Ministry of Internal Affairs and Communications (Grant number: 201605002) and Grant-in-Aid for Scientific Research (Grant number: 20H04207).

References

[1] H. Brumm and S. A. Zollinger, "The evolution of the Lombard effect: 100 years of psychoacoustic research," *Behaviour*, vol. 148, no. 11/13, pp. 1173–1198, 2011.

[2] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, 2015. DOI: 10.1109/MSP.2014.2365594.

[3] C. H. Taal and J. Jensen, "SII-based speech preprocessing for intelligibility improvement in noise," in *Proc. Interspeech 2013*, 2013, pp. 3582–3586. DOI: 10.21437/Interspeech.2013-770.

[4] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech and Language*, vol. 28, no. 4, pp. 858–872, 2014, ISSN: 0885-2308. DOI: https://doi.org/10.1016/j.csl.2013.11.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230813001101.

[5] T.-C. Zorilă and Y. Stylianou, "On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement," in *Proc. Interspeech 2014*, 2014, pp. 2050–2054. DOI: 10.21437/Interspeech.2014-466.

[6] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech 2012*, 2012, pp. 635–638. DOI: 10.21437/Interspeech.2012-197.

[7] C. Chermaz and S. King, "A Sound Engineering Approach to Near End Listening Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 1356–1360. DOI: 10.21437/Interspeech.2020-2748.

[8] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Communication*, vol. 45, no. 2, pp. 101–113, 2005, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2004.06.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639304000597.

[9] T. Ngo, R. Kubo, and M. Akagi, "Increasing speech intelligibility and naturalness in noise based on concepts of modulation spectrum and modulation transfer function," *Speech Communication*, vol. 135, pp. 11–24, 2021, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2021.09.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016763932100100X.

[10] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech 2011*, 2011, pp. 345–348. DOI: 10.21437/Interspeech.2011-126.

[11] V. Aubanel and M. Cooke, "Information-preserving temporal reallocation of speech in the presence of fluctuating maskers," in *Proc. Interspeech 2013*, 2013, pp. 3592–3596. DOI: 10.21437/Interspeech.2013-772.

[12] H. Noh and D.-H. Lee, "How does speaking clearly influence acoustic measures? a speech clarity study using long-term average speech spectra in korean language," *Clinical and Experimental Otorhinolaryngology*, vol. 5, pp. 68–73, 2012.

[13] C. Kashimada, K. Ogita, T. Ishikawa, H. Hasegawa, and M. Ayama, "Effects of voice training on subjective evaluation of voice quality," *The Journal of The Institute of Image Information and Television Engineers*, vol. 63, no. 12, pp. 1818–1823, 2009. DOI: 10.3169/itej.63.1818.

[14] M. Kobayashi and M. Akagi, "Intelligibility of announcer's speech in noisy environments," *IEICE Technical Report*, vol. 119, pp. 95–99, 2020.

[15] T. Vu Ho, M. Kobayashi, and M. Akagi, "Speak Like a Professional: Increasing Speech Intelligibility by Mimicking Professional Announcer Voice with Voice Conversion," in *Proc. Interspeech 2022*, 2022, pp. 171–175. DOI: 10.21437/Interspeech.2022-124.

[16] Y. A. Li, A. Zare, and N. Mesgarani, "StarGANv2-VC: A Diverse, Unsupervised, Non-Parallel Framework for Natural-Sounding Voice Conversion," in *Proc. Interspeech 2021*, 2021, pp. 1349–1353. DOI: 10.21437/Interspeech.2021-319.

[17] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203. DOI: 10.1109/ICASSP40776.2020.9053795.

[18] A. T. R. I. International, *Digital voice database*, http://www.atr-p.com/.

[19] P. Manuel, *Python implementation of STOI*, https://github.com/mpariente/pystoi.

[20] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994. DOI: 10.1044/jshr.3704.769.

[21] H. M. Hanson and E. S. Chuang, "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *The Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 1064–1077, 1999. DOI: 10.1121/1.427116.

[22] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. C. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice," *Journal of Speech, Language, and Hearing Research*, vol. 38, no. 6, pp. 1212–1223, 1995. DOI: 10.1044/jshr.3806.1212.

[23] R. Fraile and J. I. Godino-Llorente, "Cepstral peak prominence: A comprehensive analysis," *Biomedical Signal Processing and Control*, vol. 14, pp. 42–54, 2014, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2014.07.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809414000986.