IEICE TRANS. INF. & SYST., VOL.E97–D, NO.12 DECEMBER 2014

LETTER

Muffled and Brisk Speech Evaluation with Criterion Based on Temporal Differentiation of Vocal Tract Area Function

Masanori MORISE[†a)], ***Member***, **Satoshi TSUZUKI**[†], ***Nonmember***, **Hideki BANNO**[††], ***and*** **Kenji OZAWA**[†], ***Members***

SUMMARY This research deals with muffled speech as the evaluation target and introduces a criterion for evaluating the auditory impression in muffled speech. It focuses on the vocal tract area function (VTAF) to evaluate the auditory impression, and the criterion uses temporal differentiation of this function to track the temporal variation of the shape of the mouth. The experimental results indicate that the proposed criterion can be used to evaluate the auditory impression as well as the subjective impression.

key words: *speech analysis, vocal tract area function, auditory impression, speech evaluation*

1. Introduction

Speech synthesis and voice conversion are important research topics, and recent systems can synthesize speech that is as natural as the input speech. However, although the sound quality of speech synthesized with voice conversion is high enough to communicate linguistic information, it is difficult to obtain speech with a good auditory impression. In particular, muffled speech is a significant problem affecting voice conversion and text-to-speech systems. For this reason, it would be useful to develop a means of measuring and improving the auditory impression for speech synthesis.

In this study, we introduced a criterion and tried to evaluate the auditory impression. Muffled speech is assumed to be produced by the mouth in a certain shape that undergoes variations for articulation. The criterion therefore uses a vocal tract area function (VTAF) in order to approximate the shape of the mouth. The conventional VTAF estimation method requires magnetic resonance imaging (MRI) [1]. In addition, several methods to directly estimate the VTAF from speech waveforms have been proposed [2], [3]. In our research, we used these prior methods to estimate the VTAF and tried to evaluate the auditory impression by using the VTAF-based criterion.

2. Definition of Auditory Impression

We define the auditory impression, *brisk* speech, as the antonym of muffled speech. Here, brisk only refers to the auditory impression, and it differs from speech that has a high level of articulation. We attempted to evaluate auditory impression as well as the subjective impressions they convey and develop a criterion for evaluating them.

Speech articulation (sound and syllable articulation) is traditionally used as one of the criteria to evaluate the sound quality of speech. However, this criterion is defined as the percentage of questions on the phonemes and syllables answered correctly and cannot be used to evaluate the auditory impression. The auditory impression used in this paper differs from speech articulation because humans have the different auditory impressions regardless of the phoneme or syllable information when they listen to two different types of speech. The purpose of the criterion is to evaluate such subjective impression.

The simplest feature for comparing the difference between two types of speech is the spectrogram. Figure 1 illustrates two spectrograms. The top one represents brisk speech, and the bottom one represents muffled speech. The uttered sentence is the five Japanese vowels /aiueo/ spoken continuously by a male speaker. The formant patterns differ between the two types of speech, although it is difficult to identify this difference from the spectrograms.

We used the *vocaltractgram* (a spectrogram-like display of a VTAF sequence) shown in Fig. 2 to easily evaluate the auditory impression. The color bar represents Log-VTAF, and 0 represents a vocal tract are of 1.0 cm^2. The vocaltractgrams clearly show that the vocal tract area of brisk

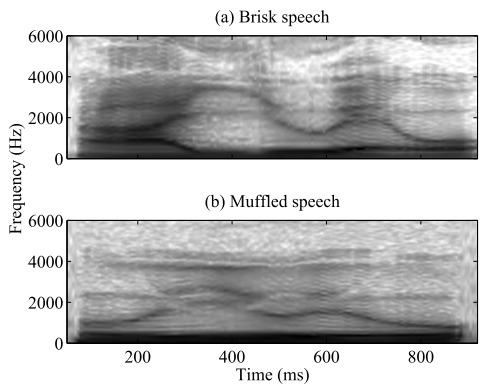

Fig. 1 Spectrograms of two types of speech. Top: Brisk speech, bottom: Muffled speech.

Manuscript received July 31, 2014.
Manuscript revised August 19, 2014.
Manuscript publicized September 17, 2014.

[†]The authors are with the Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu-shi, 400–8511 Japan.
[††]The author is with the Graduate School of Science and Technology, Meijo University, Nagoya-shi, 468–8502 Japan.
a) E-mail: mmorise@yamanashi.ac.jp

DOI: 10.1587/transinf.2014EDL8157

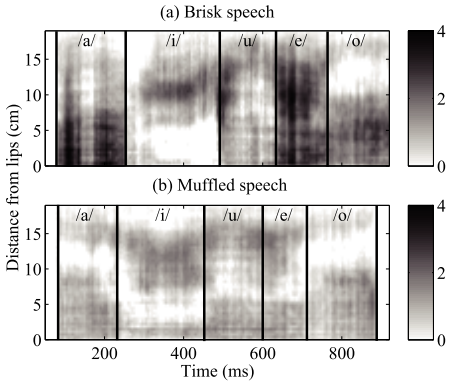

Fig. 2 Vocaltractgrams of two types of speech. Top: Brisk speech, bottom: Muffled speech. The vocal tract area of brisk speech is broader than that of muffled speech.

speech is broader than that of muffled speech, and VTAF's transition of brisk speech around the boundaries is shorter than that of muffled one. This suggests that the auditory impression depends on the vocal tract area for each vowel and temporal variation between neighboring two vowels.

3. Proposed Criterion

Here, we propose a criterion on the basis of the vocaltractgram and explain how to calculate it. A VTAF estimation method based on the sub-band line spectral pair (LSP) [3] is used to estimate the accurate VTAF. Since this method requires the accurate spectral envelope, we used TANDEM-STRAIGHT [4], [5] that fulfills the requirement.

The proposed criterion uses the temporal differentiation of the vocaltractgram $A(i,n)$ to observe the temporal variation of the shape of the mouth.

$$A(i,n) = \begin{cases} \dfrac{1+p(i,n)}{1-p(i,n)}A(i+1,n) & (i = M, M-1, \ldots, 1) \\ \text{const.} & (i = M+1) \end{cases}, \tag{1}$$

where $p(i,n)$ represents the PARCOR coefficients, n is a discrete time index, i is the index of the vocal tract area, and M is the number of stages of VTAF. In this method, $A(M+1)$ is set to 1.0 cm^2 [3]. $A'(i,n)$, the temporal derivative of $A(i,n)$, is

$$A'(i,n) = A(i,n+1) - A(i,n). \tag{2}$$

The standard deviation on the time axis in $A'(i,n)$ indicates a higher value, provided that the temporal variation of the mouth is wide. The basic criterion $s(i)$ is therefore defined as

$$s(i) = \left(\frac{1}{N-1} \sum_{n=0}^{N-2} \left(A'(i,n) - \bar{A}'(i) \right)^2 \right)^{\frac{1}{2}}, \tag{3}$$

$$\bar{A}'(i) = \frac{1}{N-1} \sum_{n=0}^{N-2} A'(i,n), \tag{4}$$

Table 1 Recording conditions.

Sentence	Five Japanese vowels /aiueo/
Microphone	AKG C414XLS
Sampling frequency	48 kHz
Quantization bit	16 bit
Number of subjects	Six males and six females
Background noise level	45 dB

where N represents the number of frames. $s(i)$ is calculated by using only the voiced section. This idea is similar to global variance (GV) [6], but with the vocaltractgram instead of the mel-cepstrum sequence. In this paper, the sum of $s(i)$ is used as the proposed criterion S for the muffled and brisk speech evaluation,

$$S = \sum_i s(i). \tag{5}$$

On the other hand, $s(i)$ is also used to discuss the effectiveness of the proposed criterion.

4. Evaluation and Discussion

We evaluated several pieces of speech recorded in a real environment and verified that the proposed criterion could appropriately evaluate brisk and muffled speech as well as the subjective impression. We also examined the relationship between the distance from the lips used for articulation and auditory impression.

4.1 Experimental Condition

Speech uttered by six males and six females was used for the evaluation. The recording conditions are shown in Table 1. Since the proposed criterion is for voiced speech, the input voice was the five Japanese vowels /aiueo/ spoken continuously. The subjects were asked to utter the vowels with the two types of speech: brisk and muffled. The subjects were not asked about the shapes made by their mouths. They were only asked to utter the two types of speech on the basis of their subjective impressions.

For TANDEM-STRAIGHT analysis, the FFT length was set to 4096 samples determined on the basis of the sampling frequency and lowest F0, and the frame shift was 5 ms. A Blackman window with a length of $2.5T_0$ was used. For VTAF estimation, the maximum value of i was set to 52. Since length of the vocal tract was $i \times 0.36$ cm in this research, and the vocal tract length was 18.72 cm.

To compare the proposed criterion with the subjective impression, a subjective evaluation was carried out by using the mean opinion score (MOS). The recorded speech was used, and nine subjects with normal hearing ability participated in the evaluation. The stimuli were randomly presented to the subjects though headphones (SENNHEISER HD650). They evaluated each stimulus on a five-grade (1: Very muffled and 5: Very brisk). A sound-proof room with an A-weighted sound pressure level (SPL) of 18 dB was used for the evaluation.

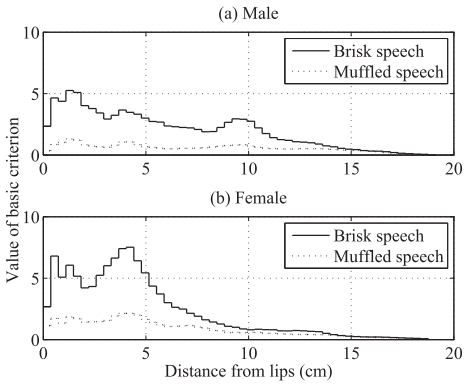

Fig. 3 Average of the basic criteria for brisk and muffled speech.

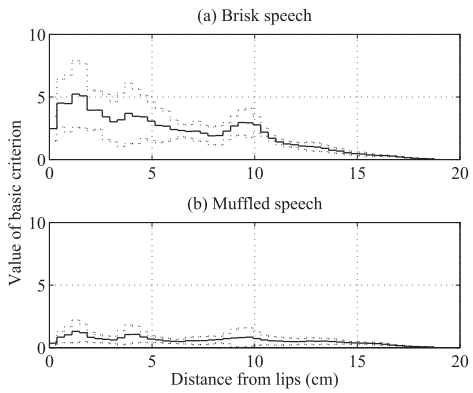

Fig. 4 Average (solid line) and standard deviation (dotted line) of the basic criteria. This figure plots the results of the male subjects.

4.2 Difference between the Criteria Calculated from Muffled and Brisk Speech

Figure 3 illustrates the average value of the basic criterion $s(i)$. The horizontal axis represents the distance from the lips, and the vertical axis represents the value of the basic criterion. This figure indicates that brisk speech (the solid line) was higher than muffled speech (the dotted line) regardless of the gender. Figure 4 illustrates the averages and standard deviations of the basic criterion. This figure represents the results of the males (the same trend was observed in the results of females). The standard deviation of brisk speech was higher than that of muffled speech. This indicated that brisk speech depends on the subject and muffled speech does not.

Figure 5 illustrates difference between the muffled and brisk speech based on the proposed criterion S of each subject. The horizontal axis represents the value of the proposed criterion, and error bar represents the 95% confidence intervals. In Fig. 5, the p-value between the two type of speech was below 0.001. This result showed that the proposed criterion could evaluate muffled speech even if there were individual differences even if brisk speech depends on the subject.

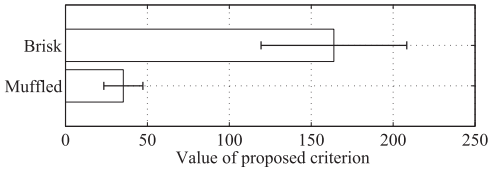

Fig. 5 Difference between brisk and muffled speech determined by the proposed criterion.

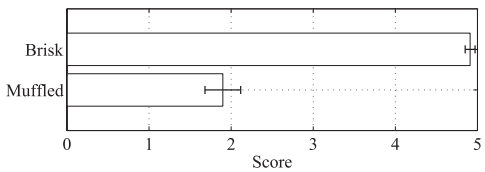

Fig. 6 Difference between brisk and muffled speech: Results of the subjective evaluation.

4.3 Subjective Impression by the Subjective Evaluation

The results of the subjective evaluation are illustrated in Fig. 6. The error bars represent 95% confidence intervals. In Fig. 6, the p-value between the two type of speech was also below 0.001 along with Fig. 5. These results show that the proposed criterion can evaluate brisk and muffled speech in accordance with the subjective impression.

4.4 Discussion

The results clearly showed that the proposed criterion can evaluate brisk and muffled speech, and the difference between these two types of speech can be observed near the lips. The result that the standard deviation of brisk speech was larger than that of muffled speech indicated that the personality of each person influenced how the mouth was controlled to make the shapes for expressing brisk speech. However, the proposed criterion could evaluate the difference between the two types of speech regardless of the personality of the speaker.

The results also indicated the difference between males and females, which is a difference between vocal tract lengths. Vocal tract length normalization [7] would be able to remove this gender difference. Moreover, voice morphing [8] enables us to convert muffled speech into brisk speech. Thus, the development of a voice conversion incorporating the proposed criterion will be an important topic of study.

5. Conclusion

A criterion based on temporal differentiation of the VTAF was used to evaluate the auditory impression in muffled and brisk speech. A temporally stable VTAF is calculated from a combination of the TANDEM-STRAIGHT spectrum and LSP-based VTAF estimation, and it enables us to evaluate

the auditory impression. The results of objective and subjective tests clearly indicated that the proposed criterion can evaluate the auditory impression in accordance with the subjective impression.

The next step in this research will be to use the proposed criterion to improve voice conversion, which means conversion from muffled to brisk speech. We will try to develop a conversion method for speech including not only vowels but also consonants.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 24300073 and 26540087 and the Research Institute of Electrical Communication, Tohoku University (H25/A08).

References

[1] B.H. Story, I.R. Titze, and E.A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," J. Acoust. Soc. Am., vol.100, no.1, pp.537–554, 1996.

[2] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," IEEE Trans. Audio Electro., vol.AU-21, no.5, pp.417–427, 1973.

[3] A. Arakawa, Y. Uchimura, H. Banno, F. Itakura, and H. Kawahara, "High quality voice manipulation method based on the vocal tract area function obtained from sub-band LSP of straight spectrum," ICASSP2010, pp.4834–4837, 2010.

[4] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," SADHANA – Academy Proceedings in Engineering Sciences, vol.36, no.5, pp.713–728, 2011.

[5] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," ICASSP2008, pp.3933–3936, 2008.

[6] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.816–824, May 2007.

[7] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," ICASSP96, vol.1, pp.346–348, 1996.

[8] H. Kawahara, R. Nisimura, T. Irino, and M. Morise, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," ICASSP2009, pp.3905–3908, 2009.