Relationship Between Speakers' Physiological Structure and Acoustic Speech Signals: Data-Driven Study Based on Frequency-Wise Attentional Neural Network

Kai Li[†], Xugang Lu[‡], Masato Akagi[†], Jianwu Dang[†], Sheng Li[‡], and Masashi Unoki[†*]
[†]*School of Information Science, Japan Advanced Institute of Science and Technology, Japan*
[‡]*Advanced Speech Technology Laboratory, National Institute of Information and Communications Technology, Japan*
[†]{*kai_li, akagi, jdang, unoki*}*@jaist.ac.jp* [‡]{xugang.lu, sheng.li}@nict.go.jp

***Abstract*—Quantitatively revealing the relationship between speakers' physiological structure and acoustic speech signals by considering the properties of resonance and antiresonance can help us to extract effective speaker discriminative information (SDI) from speech signals. The conventional quantification method based on F-ratio only considers the power of acoustic speech in each frequency band independently. We propose a novel frequency-wise attentional neural network to learn the nonlinear combined effect of the frequency components on speaker identity. The learned results indicate that antiresonance frequency induced by the nasal cavity is another essential factor for speaker discrimination that the F-ratio method could not reveal. To further evaluate our findings, we designed a non-uniform subband processing strategy based on the learned results for speaker feature extraction and did automatic speaker verification (ASV). The ASV results confirmed that further emphasizing the spectral structure around the antiresonance frequency region can enhance speaker discrimination.**

***Index Terms*—physiological feature, non-uniform filterbank, frequency-wise attention, data-driven feature**

I. Introduction

The resonance and antiresonance properties of a speaker's vocal tract are closely related to the physiological structure of their speech organs and include a lot of speaker discriminative information (SDI) [1], [2]. Therefore, quantifying the effect of frequency components on the speaker identity by considering the properties of resonance and antiresonance can help us to understand the relationship between speakers' physiological structure and acoustic speech signals and extract reliable SDI.

Previous research showed that the diversity of speech organs (e.g., the glottis [3], nasal cavity [4]–[6], piriform fossa cavities [7], [8], and vocal tract length [9]) non-uniformly provides speaker-dependent information to different frequency components in the acoustic spectrum. This SDI is encoded in the resonances and antiresonances of the speech spectrum.

*Corresponding author. This work was supported by JSPS-NSFC Bilateral Joint Research Projects/Seminars (JSJSBP120197416), a Grant-in-Aid for Scientific Research (20H04207), the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233), and the KDDI Foundation (Research Grant Program).

Many acoustic features have been successfully used for speaker recognition based on an elaborate non-uniform filterbank (NUF). These features emphasize the spectral structure around frequency regions with high speaker discrimination. The Mel-scale used in the Mel-frequency cepstral coefficient (MFCC) feature [10], Bark scale [11] and ERB scale [12] has been proposed to achieve frequency warping based on the frequency resolution of human hearing. Based on these auditory perception-inspired feature extraction methods, robust automatic speech recognition (ASR) could be achieved. However, feature extraction in ASR is for removing or normalizing SDI. Therefore, those feature extraction methods are not suitable for extracting SDI for tasks that require speaker identity information [1].

Recently, data-driven-based methods of exploring SDI have been getting a lot of attention. For example, Gaussian functions have been constructed (instead of triangular Mel-scale filterbanks) as a pseudo-filterbank layer to obtain learnable filterbanks for feature extraction in the automatic speaker verification (ASV) task [13]. Furthermore, the SincNet proposed in [14] was used to obtain superior band-pass filters in the first layer of NN with raw time-domain waveform as input. However, these methods could not provide insights into how those extracted features are connected to the physiological structure of vocal tracts, which is the physical foundation for discriminating speakers. Therefore, we aim to figure out where SDI is encoded in acoustic speech and reveal its connections to the physiological structure of the vocal tract.

Given the strong connection between the physiological structure of the vocal tract and acoustic spectral structure, it is logical to measure the relevance between them. Rather than figuring out the relevance based on analytical approaches from case studies [3]–[9], the statistical methods based on analyzing a large speaker data corpus have been proposed. For example, Fisher's F-ratio [15] is a statistical method to measure the discriminative ability of a feature for a given recognition task. This method was employed to measure the relationship between frequency components and speaker individuality (as

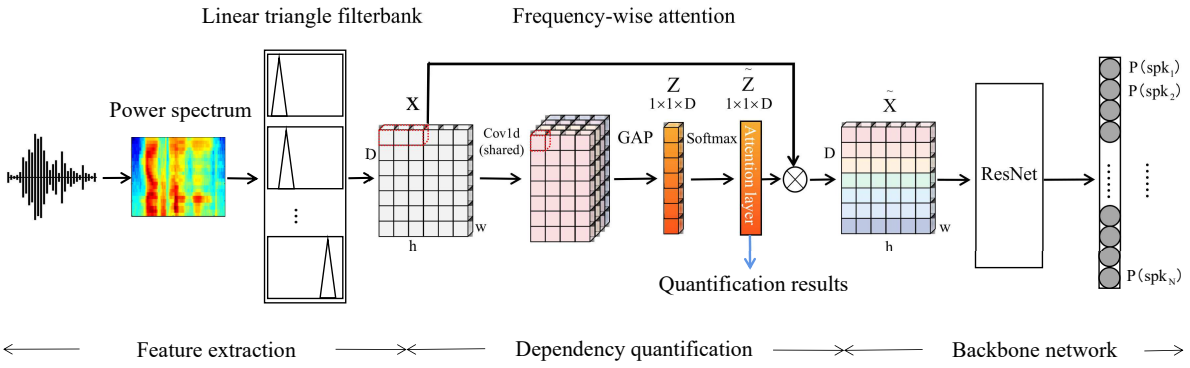

Fig. 1. Proposed residual network architecture augmented with frequency-wise attention to learn dependencies between frequency components and speaker individuality

in [1]). Then, a speaker-dependent acoustic feature, named non-uniform filterbank cepstral coefficients (NUFCC), was proposed to improve speaker identification performance. A similar method was used to improve the accuracy of speech emotion recognition [16] and replay attack detection [17].

However, the F-ratio-based quantification method only uses the statistical mean for the variance estimation with a simple single-mode model (e.g., single-mode Gaussian distribution model assumption). In addition, the F-ratio is estimated from each frequency band independently based on a power spectrum feature. It cannot consider the combined effects of each frequency component on speaker discrimination and may ignore the spectral structure around the antiresonance region because antiresonance usually contributes to a spectral valley with low power energy. The combined effects of this spectral valley (antiresonance) with spectral peaks (resonances) could encode SDI in a delicate spectral structure, which is better to explore for SDI extraction.

Therefore, we propose a novel data-driven quantification method to make up for the deficiencies of the F-ratio. Inspired by the channel-wise attention model proposed for image recognition [18], the proposed method combines a frequency-wise attention architecture with a residual network (ResNet) to learn the nonlinear combined effect of the frequency components on the speaker identity from acoustic data. Based on the attentional neural model, it is supposed that the importance of frequency components to SDI extraction could be explicitly measured. Moreover, we conducted qualitative and quantitative evaluations to check whether or not the learned importance measurement of frequency components could fit the knowledge derived from the physiological study of speech production and improve the performance of the speaker-related tasks. By emphasizing the spectral structures around those revealed important frequency regions, it is expected that the performance of ASV will be improved.

II. Proposed Quantification Method

Inspired by attention modeling in pattern recognition tasks, we designed a frequency-wise attention model to capture the importance of each frequency component when the recognition task is designed as speaker recognition. Before introducing our proposed method, first, we briefly review the F-ratio-based method of estimating the importance of each frequency component in speaker discrimination.

A. F-ratio-based measurement for speaker discrimination

Given the input acoustic features for speaker discrimination, the F-ratio is defined as:

$$\text{F-ratio} = \frac{\frac{1}{M}\sum_{i=1}^{M}(u_i - u)^2}{\frac{1}{M\cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N}(x_i^j - u_i)^2}, \tag{1}$$

where x_i^j is the acoustic feature variable (subband energy is used in this study) of the jth speech frame of speaker i with $j = 1, 2, ..., N$, and $i = 1, 2, ..., M$, and u_i and u are variables that represent the subband energy averages for speaker i and for all speakers, respectively, which are defined as:

$$u_i = \frac{1}{N}\sum_{j=1}^{N} x_i^j; \qquad u = \frac{1}{M\cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N} x_i^j. \tag{2}$$

Equation (1) is the ratio between the inter-speaker variance and intra-speaker variance of speech power in a given frequency band. A larger value obtained in a frequency band means that more speaker information is encoded in that band.

In Eq. (1), the discrimination measurement that uses F-ratio is based on a single-mode Gaussian distribution assumption of the subband power energy variable. It is possible that the distribution could be multi-mode with a mixture of distributions. In addition, the frequency importance is calculated in each frequency band independently. Therefore, it cannot reflect the nonlinear and joint relationship among different frequency bands. Particularly, the F-ratio-based method cannot explore speaker-specific information encoded by antiresonance because antiresonance usually contributes SDI to spectral valleys with low power energy.

B. Frequency-wise attention model

Deep neural network (DNN)-based models have been successfully used for ASV. Due to the strong capacity in speaker discriminative feature extraction, the performance of ASV has been significantly improved. However, as a black box modeling method in DNN, it is difficult to understand which acoustic features are specifically relevant to speaker discrimination. Unlike most studies, we obtain information about which frequency components are important for speaker discrimination by explicitly inserting a frequency-wise attention module in a DNN-based speaker recognition task. Our method consists of a frequency-wise attention architecture and a simple ResNet, which is illustrated in Fig. 1, to learn the importance of each frequency band. Motivated by channel-wise attention in image recognition [18], we designed the frequency-wise attention module to map the input feature $\mathbf{X}$ to weighted feature $\widetilde{\mathbf{X}}$, and $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_D}]$, where $\mathbf{x}_i \in \mathbb{R}^{w \times h}$ represents the i-th frequency band of input feature $\mathbf{X}$, h is the frame index, $w = 1$, D is the number of frequency components. Specifically, convolution operations are first carried out in $\mathbf{x}_i$ along the time axis using a shared one-dimensional convolution layer. We then apply global average pooling (GAP) for each channel to obtain the channel feature $\mathbf{Z}$:

$$\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_D], \tag{3}$$

where $\mathbf{z}_i$ represents the feature of the i-th frequency component. The importance of each frequency component $\widetilde{\mathbf{Z}}$ (attention layer) can be learned after using the softmax function. The weighted feature map is calculated as follows:

$$\widetilde{\mathbf{X}} = \widetilde{\mathbf{Z}} \otimes \mathbf{X}, \tag{4}$$

where $\otimes$ represents the outer product of vectors. The ResNet includes three one-dimensional convolutional layers combined with three residual blocks to generate a segment-level feature for each utterance. Finally, we use the cross-entropy loss as an objective during the optimization of the entire network. A detailed description of this ResNet can be found in our previous study [19].

C. Qualitative and quantitative evaluations

There are several ways to evaluate the learned quantification results. For example, whether the dependency of each frequency component can explain our intuitive knowledge from the speech production aspect (qualitative evaluations). Or whether emphasizing each frequency component based on the learned values can help to improve speaker individuality discrimination and, hence, improve the performance of speaker-related tasks (quantitative evaluations). In this paper, the learned result will be evaluated by using these two evaluation methods.

1) Qualitative evaluation: For qualitative evaluation, the physiological properties of the vocal tract are essential for explaining the quantification results. For example, the glottis is an important articulator to modulate the air input from the lung. The vibration frequency of a normal adult glottis ranges between 60 and 400 Hz due to the differences in glottis length and stiffness among different speakers [3]. The nasal cavity is the largest side branch within the vocal tract. The nasal cavity with the sinuses demonstrates significant SDI from 1 kHz to 2 kHz when producing nasal and nasalized sounds [4]–[6].

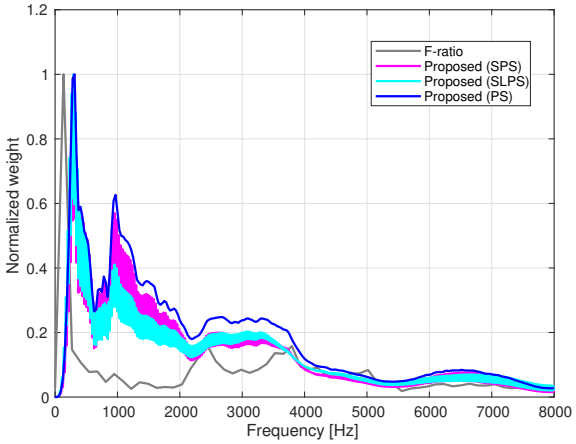

Fig. 2. Comparison of quantification results from using F-ratio-based and proposed quantification methods. Three different features were used as front-end input of proposed architecture.

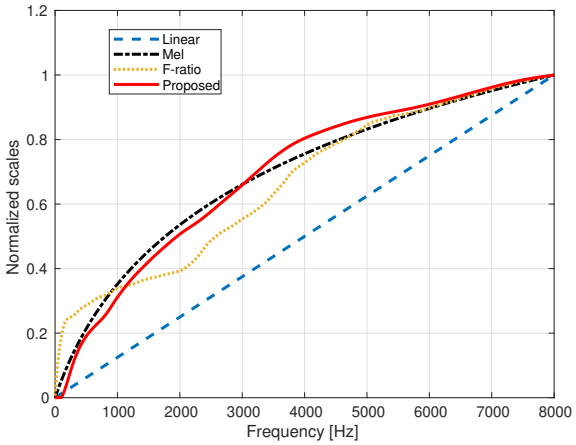

Fig. 3. Frequency warping for linear, Mel, F-ratio, and proposed scale.

2) Quantitative evaluation: For quantitative evaluation, the learned importance weightings are used to extract the NUFCC, and they are used to an i-vector-based ASV system to examine whether an ASV performance has improved. Specifically, to emphasize the importance of frequency regions with relatively high quantification scores, the distribution density of the triangular band-pass filters is assigned to be directly proportional to the average quantification score (Q_{score}). Q_{score} is calculated by:

$$Q_{score} = \frac{\sum_{i=1}^{N} \widetilde{\mathbf{Z}}_i}{N}, \tag{5}$$

where N is the number of utterances, $\widetilde{\mathbf{Z}}_i$ is the quantification score of ith utterance. The steps for designing an NUF are as

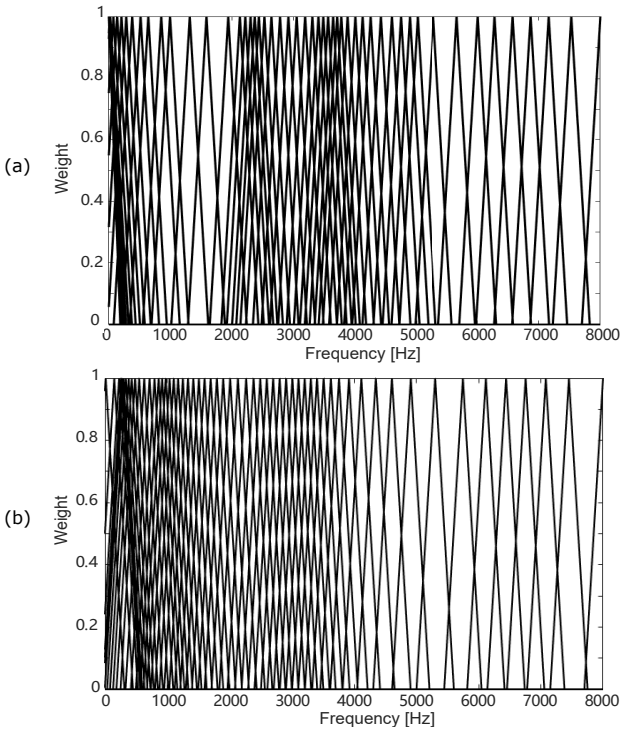

Fig. 4. Comparison of NUFs designed with F-ratio-based method (a) and proposed method (b). Number of filters was 60, and bandwidth of each sub-band filter was fixed.

follows:

- calculate the weight k based on the Q_{score}, $k = fs/(2 \times Sum(Q_{score}))$, where fs is the sampling rate,
- calculate the cumulative Sum of weighted Q_{score}, $Sum = Cumsum(k \times Q_{score})$,
- fit the curve of the mapping frequency from the linear scale to the adaptive scale by cubic spline interpolation,
- calculate the center frequency of the triangular band-pass filters $C(i)$ based on the fitting curve, and
- design an NUF with the same bandwidth.

The designed NUF is used instead of the Mel-filterbank to obtain a novel NUFCC in the MFCC-extraction process.

The i-vector technique proposed by Dehak et al. [20], [21] is a commonly used baseline system in speaker recognition. It establishes a low-dimensional total-variability space that simultaneously models speaker and channel variability. We designed an i-vector system based on the proposed NUFCC and implemented it using Kaldi [22] to evaluate the effectiveness of the proposed quantification method.

III. EXPERIMENTS

A. Database

The Japanese versatile speech corpus [23] consists of audios from 100 native Japanese speakers. The database was recorded in a clean environment at a 24-kHz sampling rate. To train our model, we selected a set of 9,997 sentences from 100 speakers to learn the nonlinear combined effect of the frequency components on speaker identity. All the sentences were downsampled from 24 kHz to 16 kHz. The average length of each utterance was 7.92 s, and the total length of the speech data was 21.86 hrs. In i-vector-based ASV, the same speech data introduced above was divided into training (70 speakers) and testing (30 speakers) sets.

TABLE I
RESULTS OF OUR DESIGNED I-VECTOR-BASED ASV SYSTEMS IN TERMS OF EER AND MINDCF BASED ON A JAPANESE DATABASES.

Acoustic feature	EER (%)	minDCF (0.01)
UFCC	3.092	0.417
MFCC	2.977	0.363
F-ratio-based NUFCC	2.084	0.215
Proposed NUFCC (SLPS)	1.800	0.219
Proposed NUFCC (PS)	1.698	0.236
Proposed NUFCC (SPS)	**1.597**	**0.206**

B. Experimental conditions

We used power spectrum (PS), subband power spectrum (SPS), and subband log power spectrum (SLPS) as the front-end input. The extraction of these three features was without frequency warping operations. The filterbank used for SPS and SLPS features extraction was a triangular band-pass filter with a linear frequency scale, and the dimension was set to 512. In the frequency-wise attention module shown in Fig. 1, the kernel size of the one-dimensional convolution layer is (5×1), and the number of output channels is 64. The dimension of the attention layer corresponds to the number of frequency components that were set to 512.

For i-vector-based ASV, the UBM and i-vector extractor were trained on the training set, and 30,000 test pairs, including half positive trials and half negative trials, were randomly generated from the testing set. The Gaussian mixture number of UBM was set to 128, and the dimension of the i-vector was set to 300. We used the equal error rate (EER) and minimum decision cost function (minDCF) with P_{target} = 0.01 as the evaluation metrics of the ASV [24].

IV. RESULTS AND DISCUSSION

A. Quantification results using proposed method

Figure 2 shows the speaker discriminative abilities of each frequency component quantified using the F-ratio-based quantification method and our quantification method. The comments in the parentheses refer to the front-end input feature types for the training of our method. We compare the two methods by showing all the results using normalization with values ranging from 0 to 1. The quantification results show the distribution of SDI in the frequency domain was non-uniform and most of the discriminative information concentrated in the low-frequency region. Using our method with different input features, we could obtain consistent results with peaks and valleys located in similar frequency regions on the curves. Moreover, the quantification results with PS as an input feature had fewer fluctuations than others.

Figure 3 illustrates the normalized plot of different frequency warping scales to compare our method with other

methods. We can observe that the frequency warping based on data-driven methods has a high frequency resolution in the low-frequency regions (below 400 Hz), which is discriminative information expected from the glottis based on knowledge from [3]. Compared to the F-ratio-based quantification method, the normalized scale from our method (red-solid curve) indicates a higher frequency resolution from 1 kHz to 2 kHz. Based on [5] and [6], this peak is possibly related to the antiresonance contributed by the nasal cavity and sinuses, which is another essential factor for speaker discrimination that the F-ratio method could not reveal.

B. Effectiveness of NUFCC designed with proposed method for ASV

Based on different frequency warping scales, NUFs can be designed using the steps described in Section II-C. Two examples of the designed NUF are depicted in Fig. 4. The specially designed NUF can extract the speaker features for ASV. The ASV results in Table I indicate that acoustic feature extraction using an NUF can substantially improve speaker discrimination abilities. In addition, NUFCC extraction with our method can perform better than the F-ratio-based quantification method in both EER and minDCF. This also indicates that the quantification results from using our method can capture more speaker discriminative factors, such as the relationships among different frequency bands. The NUFCC feature designed with our quantification method using SPS as input decreases the EER from 2.084% (F-ratio-based method) to 1.597, resulting in a relative improvement of 23.4%.

V. CONCLUSION

We quantified the nonlinear combined effect of frequency components on speaker identity. A frequency-wise attention structure combined with a ResNet was designed to learn the importance of different frequency bands by considering resonance and antiresonance. The quantification results with our method using three input features consistently indicated that SDI is non-uniformly distributed in the frequency domain and most of the discriminative information is concentrated in the low-frequency region. In addition, the quantification results from using our method indicated that the antiresonance frequency induced by the nasal cavity from 1 kHz to 2 kHz is another essential factor for speaker discrimination that the F-ratio method could not reveal. To further evaluate our findings, we designed a non-uniform subband processing strategy based on the quantification results using our method for speaker feature extraction and did ASV. Finally, compared with the NUFCC designed with the F-ratio-based method, NUFCC designed with our method achieved 23.4% relative improvement in EER. These results also confirmed that further emphasizing the spectral structure around the antiresonance frequency region could enhance speaker discrimination.

REFERENCES

[1] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," Speech Communication, vol. 50, no. 4, pp. 312–322, 2008.

[2] K.-Y. Leung, M.-W. Mak, M.-H. Siu, and S.-Y. Kung, "Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification," Speech Communication, vol. 48, no. 1, pp. 71–84, 2006.

[3] B. S. Atal, "Automatic recognition of speakers from their voices," Proceedings of the IEEE, vol. 64, no. 4, pp. 460–475, 1976.

[4] G. Fant, Acoustic theory of speech production, Number 2. Walter de Gruyter, 1970.

[5] J. Dang, K. Honda, and H. Suzuki, "Morphological and acoustical analysis of the nasal and the paranasal cavities," J. Acousti. Soc. Am., vol. 96, no. 4, pp. 2088–2100, 1994.

[6] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation," J. Acousti. Soc. Am., vol. 100, no. 5, pp. 3374–3383, 1996.

[7] J. Dang and K. Honda, "An improved vocal tract model of vowel production implementing piriform resonance and transvelar nasal coupling," in Proc. ICSLP'96, pp.965-968, Philadelphia, USA, 1996.

[8] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans," J. Acousti. Soc. Am., vol. 101, no. 1, pp. 456–465, 1997.

[9] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in Proc. IEEE-ASRU, pp. 559–564, 2011.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[11] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," J. Acousti. Soc. Am., vol. 68, no. 5, pp. 1523–1525, 1980.

[12] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," Acta Acustica united with Acustica, vol. 82, no. 2, pp. 335–345, 1996.

[13] H. Seki, K. Yamamoto, and S. Nakagawa, "A deep neural network integrated with filterbank learning for speech recognition," in Proc. IEEE-ICASSP, pp. 5480–5484, 2017.

[14] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in Proc. IEEE-SLT, pp. 1021–1028, 2018.

[15] A. R. Webb, Statistical pattern recognition, John Wiley Sons, 2003.

[16] Y. Zhou, Y. Sun, J. Li, J. Zhang, and Y. Yan, "Physiologically-inspired feature extraction for emotion recognition," in Proc. INTERSPEECH, 2009.

[17] S. Hyon, J. Dang, H. Feng, H. Wang, and K. Honda, "Detection of speaker individual information using a phoneme effect suppression method," Speech Communication, vol. 57, pp. 87–100, 2014.

[18] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in Proc. CVPR, pp. 5659–5667, 2017.

[19] K. Li, M. Akagi, and Y. Wu, "Segment-level effects of gender, nationality and emotion information on text-independent speaker verification," in Proc. INTERSPEECH, 2020.

[20] N. Dehak, R. Dehak, P. Kenny, N. Br ummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in Proc. INTERSPEECH, 2009.

[21] S. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification.," in Proc. Odyssey, p. 16, 2010.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The kaldi speech recognition toolkit," in Proc. IEEE-ASRU, 2011.

[23] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," arXiv preprint arXiv:1908.06248, 2019.

[24] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," arXiv preprint arXiv:1304.2865, 2013.