Exploiting Phase-based Features for Whisper *vs.* Speech Classification

Nirmesh J. Shah[1], M. Ali Basha Shaik[2], Periyasamy P.[2], Hemant A. Patil[1] and Vikram Vij[2]
[1]*Speech Research Lab, DA-IICT, Gandhinagar, India.*
[2]*Samsung R&D Institute, Bangalore (SRI-B), India.*
Email: {nirmesh88_shah, hemant_patil}@daiict.ac.in; {m.shaik, periyasamy.p, vikram.v}@samsung.com

***Abstract*—Performance of Voice Assistant (VA) deteriorates notably when tested on the whispered speech. Hence, separate systems are being developed for the whisper. To that effect, detecting the incoming signal as to whether it is a whisper or a speech (especially with a low latency) in the noisy environments is more desirable from the model switching point of view. We propose to exploit high resolution property of group delay spectrum (GDSPEC) to capture characteristic of excitation source (voiced vs. unvoiced) and formant shift for the early robust detection of the whispered speech. The effectiveness of the proposed feature set is investigated across different deep learning-based classifiers using three databases, namely, wTIMIT, CHAINS, and in-house database of Samsung. We obtain 3.4%, and 5.05% relative improvement in classification accuracy with the SEPC+GDSPEC compared to the individual SPEC, and GDSPEC features, respectively. Furthermore, robustness is shown in the presence of state-of-the-art noises (from the MUSAN database) for different SNR levels. Mathematical intuitions behind robustness of group delay functions are also presented. Finally, the frame-level decision was combined to predict whispered speech at an utterance-level based on the majority rule for different lengths of the speech segments.**

***Index Terms*—Whisper Detection, Spectrum, Group Delay Function, DNN, CNN, Xception**

I. Introduction

Speech technologies have made remarkable progress with the advent of Voice Assistant (VA) since the last decade [1]. Present speech systems are mostly designed for normal speech. However, people prefer to whisper for specific applications, such as private conversation in public places, conversation in library, hospital or a meeting room [2]. Significant reduction in the performance has been observed when whispered speech is directly applied to the speech systems that are trained on normal speech [3], [4]. Hence, research focus has been shifted for designing various separate speech systems for the whispered speech due to its interesting commercial applications [3]–[10]. Thus, early detection of whisper would be more valuable for the model switching in the VA. Despite significant differences in the formant frequencies, formant bandwidths and its shifts (from the speech production-perception viewpoint) in normal *vs.* the whispered speech [2], [11]–[17], these differences cannot be applied directly for detection of whispered speech [18]. This is primarily due to the fact that there are no standard reference values for these differences [18]. For example, differences in formant frequencies, and their shifts for a particular speaker can easily be masked by the shifts due to different speakers. In addition, different linguistic content spoken by different speakers also affect these differences. Hence, learning a unique representation that can discriminate speech and the whisper across speakers, channel variations is a complex task.

In this paper, we propose frame-level classifier for the robust detection of the whispered speech. This is primarily achieved via developing whisper *vs.* speech classifier. The goal is to predict whether the incoming signal to the VA is whispered or normal speech at an utterance-level as early as possible. Earlier approaches utilize spectral power ratio in a low frequency band to the high frequency band [18]. In addition, spectral information entropy (SIE) ratio-based features with Gaussian Mixture Model (GMM) classifiers have been proposed for the classification of the whispered speech from five different types of speech signals [19]. Recently, Deep Neural Network (DNN), Long Short Term Memory (LSTM) architecture with the log-filterbank energies have been proposed for the whispered detection task at the frame-level [6]. In addition, Convolutional Neural Network (CNN)-based classifier to detect whispered speech at an utterance-level in the presence of imbalance class learning [20]. However, decoding of LSTM takes more time and hence, affect the decision taking in real-time due to its dependency on the contextual neighboring frames.

Here, we propose to exploit phase spectrum-based features (i.e., high resolution group delay function) along with the magnitude spectrum-based features for the frame-level whisper detection task. Recently, phase-based features have been applied for various speech applications [21]–[24]. In addition, Fourier Transform (FT) phase-based features have been utilized to detect emotions in the whispered speech [25]. However, phase-based features that exploit high resolution property of group delay function have not been explored for the detection of the whispered speech task to the best of authors' knowledge. The key difference between whispered speech and normal speech is the complete absence of vocal fold vibrations in the whispered speech (i.e., at excitation source-level). The phase-based features are well known to capture source-excitation related information more so for whispered speech that has predominately unvoiced source [22], [26]–[28], which motivated us to explore different phase-based features for this task.The effectiveness of the proposed feature set has been shown across different classifiers and different databases in presence of noises (usually present in the home and office environment) chosen from the standard open source MUSAN database at various SNR levels [29]. Finally, the frame-level decision is combined to predict at an utterance-level (based on the majority for different lengths of the speech segments) for an early detection of whispered speech.

II. Proposed systems

A. Magnitude and Phase Spectrum-based Features

In this work, we primarily explore magnitude and phase-based features for the given task. For a given speech signal, $x(n)$, the Discrete-Time Fourier Transform (DTFT) of a speech can be represented in polar form as [30]:

$$X(e^{j\omega}) = |X(e^{j\omega})|e^{j\angle X(e^{j\omega})}, \tag{1}$$

where $|X(e^{j\omega})|$ and $\angle X(e^{j\omega})$ are the magnitude and phase spectrum, respectively, for a speech segment. Extracting discriminative features from the phase spectrum is a challenging task, since phase is not continuous in the frequency-domain due to phase wrapping (phenomenon reflecting trigonometric properties of *arctan* function). Hence, cosine operation is applied on the unwrapped phase spectrum (i.e., Cosine Phase Spectrum (CPSPEC)) [23]. However, phase unwrapping is itself challenging. To alleviate this, group delay function-based technique is applied to extract meaningful features from the phase spectrum. Group delay function is defined as the negative derivative of the phase spectrum w.r.t. the ω, which is given by [22], [31]:

$$\begin{aligned}\tau(e^{j\omega}) &= -\frac{d(\angle X(e^{j\omega}))}{d\omega} \\ &= \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + X_I(e^{j\omega})Y_I(e^{j\omega})}{|X(e^{j\omega})|^2},\end{aligned} \tag{2}$$

where $X(e^{j\omega})$ and $Y(e^{j\omega})$ are the DTFT of $x(n)$ and $y(n) = nx(n)$, respectively. Using digital resonator design and due to eq. (2), $\tau(e^{j\omega})$ for cascade resonator becomes additive and thus, high resolution in frequency-domain.

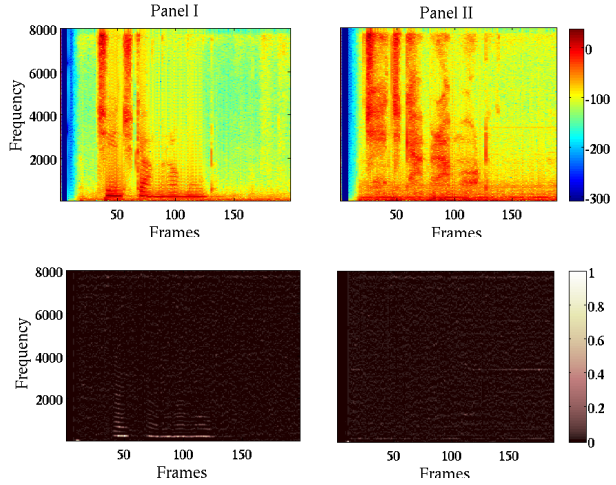

Fig. 1: An example of SPEC and GDSPEC feature for an utterance "Change Camera Mode" for normal (Panel I) and whispered speech (Panel II), respectively.

We exploit this property for whisper detection task, in particular, to capture phase characteristics of source and formant shifts. We call $\tau(e^{j\omega})$ as the group delay function-based phase spectrum (i.e., GDSPEC). SPEC and GDSPEC feature for normal and whispered speech is shown in Fig. 1. In both SPEC (top row) and GDSPEC (bottom row) difference between normal and corresponding whispered speech can be clearly observed.

B. Robustness of Group Delay Function

In this sub-Section, we establish the robustness of $\tau(e^{j\omega})$ under signal degradation conditions. Let $x(n)$ is a clean speech degraded by uncorrelated zero-mean, additive noise, $v(n)$ with variance $\sigma_v^2(e^{j\omega})$. Then, noisy speech, $z(n)$ is given by [22],

$$z(n) = x(n) + v(n). \tag{3}$$

Taking DTFT on both the sides, we get,

$$Z(e^{j\omega}) = X(e^{j\omega}) + V(e^{j\omega}). \tag{4}$$

Multiplying by the corresponding complex conjugates and taking statistical expectation operator, we have the power spectrum,

$$P_Z(e^{j\omega}) = P_X(e^{j\omega}) + \sigma_V^2(e^{j\omega}), \tag{5}$$

where $P_Z(e^{j\omega}) = E\{|Z(e^{j\omega})|^2\}$, $P_X(e^{j\omega}) = E\{|X(e^{j\omega})|^2\}$, and we assume that expectation of noise term is zero. From Eq. (5), we have,

$$P_Z(e^{j\omega}) = \sigma_V^2(e^{j\omega})\Big[1 + \frac{P_X(e^{j\omega})}{\sigma_V^2(e^{j\omega})}\Big]. \tag{6}$$

Taking logarithm on both sides and using Taylor series expansion of $ln(1+x)$, and ignoring the higher order terms in Taylor series, we get,

$$\begin{aligned}\ln(P_Z(e^{j\omega})) &= \ln\Big(\sigma_V^2(e^{j\omega})\Big[1 + \frac{P_X(e^{j\omega})}{\sigma_V^2(e^{j\omega})}\Big]\Big), \\ &\approx \ln(\sigma_V^2(e^{j\omega})) + \frac{P_X(e^{j\omega})}{\sigma_V^2(e^{j\omega})}\end{aligned} \tag{7}$$

Since, $P_X(e^{j\omega})$ is a periodic continuous function of ω with period $\omega = 2\pi$, we can expand it as a Fourier series, and we get,

$$\begin{aligned}\ln(P_Z(e^{j\omega})) \approx &\ln(\sigma_V^2(e^{j\omega})) \\ &+ \frac{1}{\sigma_V^2(e^{j\omega})}\Big[\frac{d_0}{2} + \sum_{k=1}^{+\infty} d_k cos\Big(\frac{2\pi}{\omega_0}\omega k\Big)\Big],\end{aligned} \tag{8}$$

where d_k's are Fourier series coefficients in the expansion of $P_X(e^{j\omega})$. Since $P_X(e^{j\omega})$ being $E\{|X(e^{j\omega})|^2\}$, it's an *even* function and hence, coefficients of sine terms are zero. Assuming additive noisy speech as *minimum* phase [31], we can express group delay functions in terms of cepstral coefficients [22], i.e.,

$$\tau_Z(e^{j\omega}) \approx \frac{1}{\sigma_V^2(e^{j\omega})} \sum_{k=1}^{+\infty} k d_k cos(k\omega). \tag{9}$$

Eq. (9) shows that the group delay function of noisy speech is *inversely* proportional to the noise power. In frequency region, where the noise power is greater than the signal power, i.e., high noise region in the power spectrum. Thus, group delay function of noisy speech preserves the peaks in $\tau_Z(e^{j\omega})$, which are known to carry formant (resonant) structures and thus,

helps in capturing characteristics of speech in the presence of noise. This has also positively reflected in the task of whisper detection under signal degradation conditions.

C. Details of Classifiers

To measure the effectiveness of the proposed signal processing-based features, DNN, CNN, and Xception network have been explored. DNN architecture contains three hidden layers with number of neurons, 128, 64, and 32, respectively, followed by batch normalization, and Rectifier Linear Unit (ReLU) activation nonlinearity. Batch normalization was found to be helpful while generalizing the proposed model across different databases. Sigmoid activation is applied at the output layer. The CNN architecture with three hidden convolutional layers followed by batch normalization, and ReLU activation function have been also used. Motivated from the recent studies [32]–[34], instead of fully-connected output layer, we also used convolutional layers at the output followed by sigmoid nonlinearity.

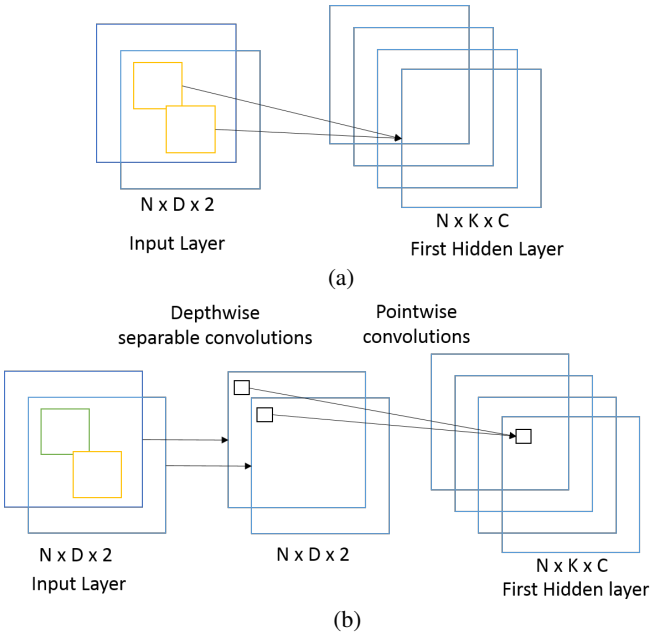

Fig. 2: Schematic representations of (a) CNN and (b) Xception architectures. After [35]. Here N: number of features in a batch, D: feature dimension, C: number of channel, K: feature dimension after convolution operation.

In conventional CNN [36], both the channels are mapped jointly with the same filter. However, cross-channel correlations between both magnitude and phase spectra may be sufficiently decoupled. Thus, depthwise separable convolution (i.e., a spatial convolution is applied independently for each channel), which is followed by pointwise convolution, which is called Xception network [35]. Furthermore, two input channels are taken one for the magnitude spectrum, and the other for the phase spectrum. The architecture of CNN, and Xception was almost the same except the fact that in Xception network, different filter is learned for each channel. However, size of the filter is the same as shown in Fig. 2. In particular, CNN architecture contains three convolutional layers with the filter size 11×11, 7×7, and 5×5. Number of channels were 32, 16, and 8. Output layer is also convolutional in both the architectures with the filter size 18x18 to match with the 2-D output for applying logistic regression. Adam optimization is used for training all the classifiers [36].

III. Experimental Results

Statistics of the databases are shown in Table I. We have used Samsung's VA (namely, Bixby) for the in-house recording. All the three databases contain data from the multiple speakers. Training was done using wTIMIT database for all the experiments. wTIMIT database was divided in three parts, namely, Train, Validation, and Eval sets. Magnitude spectrum (i.e., SPEC), cosine phase spectrum (i.e., CPSPEC), and group delay-based phase spectrum (i.e., GDSPEC) features are extracted over 25 ms Hamming window length with 10 ms frame-shift. We follow the standard classification accuracy in %, precision, recall, and F1-score for reporting our experimental results [37]. In the context of binary classification, accuracy of test is measured via F1-score [37]. F1- score is the harmonic mean of precision and recall. Best value and the worst value for F1-score is 1 and 0, respectively. Table II presents the accuracy of the DNN classifier across different features. We observe that the feature-level fusion of SPEC and GDSPEC performs better across all the databases. In particular, there is 3.4%, and 5.05% relative improvement with the proposed SPEC+GDSPEC features w.r.t. the individual SPEC, and GDSPEC features, respectively. Similar improvements have been obtained for CNN and Xception (details not presented due to space limitations). Table IV presents the performance of the proposed features across different classifiers. We observe that the proposed feature set performs consistently better across different classifiers and the databases. Since training is done on wTIMIT data, it naturally leads to the better results on wTIMIT eval test set.

TABLE I: Statistics of the Databases. Utt: Utterances

Database		Whisper		Normal	
		Duration	No. Utt.	Duration	No. Utt.
wTIMIT	Train	11:51:22	8170	11:36:51	8203
	Validation	3:58:11	2730	3:52:45	2730
	Eval	1:00:50	696	0:59:41	698
CHAINS	Eval	2:28:28	1332	2:33:08	1332
In-house (Samsung)	Eval	2:44:42	2652	2:32:42	2655

TABLE II: Classification Accuracy for Different Feature Sets

Training →	Trained on wTIMIT Train Set					
Database →	wTIMIT Eval		CHAINS		In-house	
Feature Sets ↓	Acc.	F1-Scr	Acc.	F1-Scr	Acc.	F1-Scr
SPEC	99.94	99.94	98.94	98.96	92.78	92.93
CPSPEC	99.98	99.98	99.42	99.42	94.7	94.72
GDSPEC	99.83	99.83	97.98	98.03	91.36	91.56
SPEC + CPSPEC	100	100	99.89	99.89	74.19	78.84
SPEC + GDSPEC	**99.94**	**99.94**	**99.78**	**99.79**	**95.98**	**95.9**

Acc: Accuracy, Scr: Score

We selected approximately 20 hours of noisy data (6 hours of noise, and 14 hours of music data from publicly available MUSAN database [29]) to evaluate the robustness of the

TABLE III: Performance Evaluation (% Accuracy) in the Presence of the MUSAN Database for Different SNR Levels

Classifier		Trained on wTIMIT Train Set with 0 dB SNR					
		wTIMIT Eval		CHAINS		In-house	
	Noise Level	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
DNN	Clean	**99.99**	99.99	99.92	99.92	91.79	92.07
	10 dB SNR	**99.99**	99.99	99.93	99.93	**89.6**	90.18
	0 dB SNR	99.87	99.87	**96.27**	96.22	**82.35**	84.27
	-5 dB SNR	**96.34**	96.36	**86.19**	85.1	**73.94**	77.2
CNN	Clean	**99.99**	99.99	**99.99**	99.99	**92.47**	92.67
	10 dB SNR	**99.99**	99.99	**99.99**	99.99	87.73	88.66
	0 dB SNR	**99.97**	99.97	95.21	95.06	81.49	83.7
	-5 dB SNR	95.78	95.82	84.23	83.07	72.38	76.19
Xception	Clean	**99.99**	99.99	99.97	99.97	86.03	87.21
	10 dB SNR	**99.99**	99.99	99.96	99.96	83.47	85.23
	0 dB SNR	99.85	99.85	95.55	95.43	78.4	81.38
	-5 dB SNR	94.07	94	83.59	81.74	72.37	75.62

proposed system. The noise has been randomly added to all the database for three different SNR levels of 10 dB, 0 dB, and -5 dB (i.e., severe signal degradation). The primary goal of the proposed system is to detect whisper speech in noisy (signal degradation) as well as in clean conditions. Hence, we selected proposed SPEC+GDSPEC features and trained the system on wTIMIT database in presence of 0 dB noise. The results for matched (i.e., trained and tested on 0 dB) and mismatched (i.e., trained on 0 dB and tested on 10 dB, -5 dB, and clean) conditions are presented in Table III. We observed that in the presence of noise, compared to the CNN, and Xception, on an average, DNN is performing better across the databases. Higher values of F1-score obtained across all the systems clearly indicate that both precision and recall of the classifiers are having good results [37]. We believe that with more amount of training data, and more number of hidden layers, performance of CNN and Xception can be improved further [35], [36].

TABLE IV: Classification Accuracy (in %) for SPEC+GDSPEC Features Across Different Classifiers

Nature	Trained on wTIMIT Train Set					
Database	wTIMIT Eval		CHAINS		In-house	
Classifier	Acc.	F1-score	Acc.	F1-Score	Acc.	F1-Score
DNN	99.94	99.94	99.78	99.79	95.98	95.9
CNN	**100**	100	**99.99**	99.99	**96.25**	96.2
Xception	99.99	99.99	99.98	99.98	94.28	94.27

Finally, we consider frame-level decision of each frame belonging to the given segment of the speech, and utterance was declared either whisper or normal based on majority of the frame-level decision. Utterance-level decision is taken for different amount of time from the frame-level decision. This decision was compared at an utterance-level and corresponding accuracies are shown for the Samsung's in-house database for clean and noisy cases in matched and mismatched conditions (as shown in Fig. 3). In addition, it can be observed that from 100 ms onward, utterance-level decision almost gets converged to its best value. This is possibly due to the fact that at any amount of time to predict correct decision, we need only 50% vote, and our frame-level classifiers are almost having more than 90% accuracy. Hence, for the given number of frames, getting 50% of the correct decision at the frame-level has become easier, which has clearly helped in the early detection of the whispered speech at an utterance-level. Similar results have been obtained for all the three databases (not shown due to space limitations).

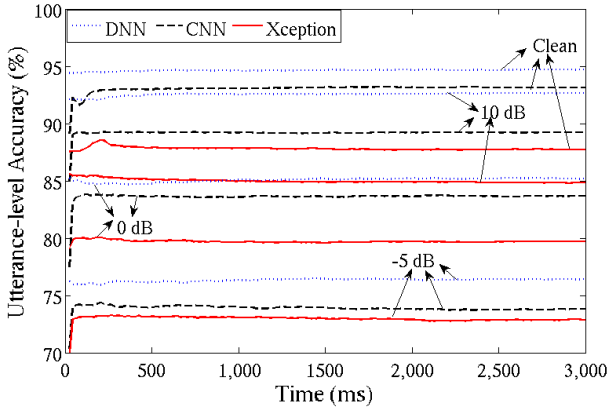

Fig. 3: Utterance-level accuracy in the noisy (signal degradation) environments for Samsung's in-house database.

IV. Summary and Conclusions

In this paper, we proposed combined magnitude and group delay-based phase spectrum (that is known to have high resolution than the traditional Fourier spectrum) for the early robust detection of the whispered speech. We presented our results on three different databases, namely, wTIMIT, CHAINS, and Samsung's in-house database. It has been observed that proposed SEPC+GDSPEC feature obtains 3.4%, and 5.05% relative improvement compared to the individual SPEC, and GDSPEC features, respectively. In addition, proposed features performed consistently better across three different classifiers, namely, DNN, CNN and Xception. Furthermore, robustness of the proposed feature has been observed in presence of state-of-the-art noises taken from the MUSAN database at different SNR levels. Our results indicate that SPEC+GDSPEC feature consistently perform better across, different databases, classifiers, and under signal degradation conditions. In addition, it has been shown that with high frame-level accuracies, it is possible to detect incoming whispered speech as early as possible at an utterance-level based on the majority. In future, we would like to evaluate our proposed features for the *imbalanced* class learning.

REFERENCES

[1] Fuliang Weng, Pongtep Angkititrakul, Elizabeth E Shriberg, Larry Heck, Stanley Peters, and John H. L. Hansen, "Conversational in-vehicle dialog systems: The past, present, and future," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 49–60, 2016.

[2] Chi Zhang and John H. L. Hansen, *Advancements in whispered speech detection for interactive/speech systems*, H. A. Patil et. al. (Eds), Signal and Acoustic Modelling for Speech and Communication Disorders, De Gruyter, vol. 5, pp. 9–32, 2018.

[3] Dorde T Grozdic and Slobodan T Jovicic, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 12, pp. 2313–2322, 2017.

[4] Qin Jin, Szu-Chen Stan Jou, and Tanja Schultz, "Whispering speaker identification," in *IEEE International Conference on Multimedia and Expo*, Beijing, China, 2007, pp. 1027–1030.

[5] Abinay Reddy Naini et al., "Formant-gaps features for speaker verification using whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 6231–6235.

[6] Zeynab Raeesy et al., "LSTM-based whisper detection," in *IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, 2018, pp. 139–144.

[7] Yaakov Chen and Doron Koren, "Whispered speech detection," Jan. 9 2018, US Patent 9,867,012.

[8] Seok Jin Hong, "Method and apparatus for recognizing whisper," Jan. 21 2016, US Patent App. 14/579,134.

[9] John S Graham, "Headset with whisper mode feature," Jan. 29 2013, US Patent 8,363,820.

[10] Marius Cotescu et al., "Voice conversion for whispered speech synthesis," *IEEE Signal Processing Letters*, vol. 27, no. 1, pp. 186–190, 2019.

[11] Nirmesh Shah, Mihir Parmar, Neil Shah, and Hemant A. Patil, "Novel MMSE DiscoGAN for cross-domain whisper-to-speech conversion," in *MLSLP Workshop*, Google Office, Hyderabad, India, 2018, pp. 1–3.

[12] Aravind Illa et al., "A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017, pp. 5075–5079.

[13] Vivien C Tartter, "What is in a whisper," *The Journal Of Acoustical Society of America (JASA)*, vol. 86, no. 5, pp. 1678–1683, 1989.

[14] Kenneth N Stevens, *Acoustic Phonetics*, MIT Press, 2000.

[15] Gokul Srinivasan, Aravind Illa, and Prasanta Kumar Ghosh, "A study on robustness of articulatory features for automatic speech recognition of neutral and whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 5936–5940.

[16] Gerard B Remijn et al., "A near-infrared spectroscopy study on cortical hemodynamic responses to normal and whispered speech in 3-to 7-year-old children," *J. of Speech, Lang., and Hearing Research*, vol. 60, no. 2, pp. 465–470, 2017.

[17] Boon Pang Lim, *Computational differences between whispered and non-whispered speech*, Ph.D. Thesis, University of Illinois at Urbana-Champaign (UIUC), USA, 2011.

[18] Stanley J Wenndt, Edward J Cupples, and Richard M Floyd, "A study on the classification of whispered and normally phonated speech," in *ICSLP*, Denver, Colorado, USA, 2002.

[19] Chi Zhang and John H. L. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2289–2292.

[20] Takanori Ashihara et al., "Neural whispered speech detection with imbalanced learning," in *INTERSPEECH*, Graz, Austria, 2019, pp. 3352–3356.

[21] Pejman Mowlaee, Rahim Saeidi, and Yannis Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[22] Hema A Murthy and Bayya Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana*, vol. 36, no. 5, pp. 745–782, 2011.

[23] Zhizheng Wu, Eng Siong Chng, and Haizhou Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *INTERSPEECH*, Portland, OR, USA, 2012, pp. 1700–1703.

[24] Meng Liu et al., "Replay attack detection using magnitude and phase information with attention-based adaptive filters," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 6201–6205.

[25] Jun Deng et al., "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2016.

[26] K Sri Rama Murty and Bayya Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE signal processing letters*, vol. 13, no. 1, pp. 52–55, 2005.

[27] K Sreenivasa Rao, SR Mahadeva Prasanna, and Bayya Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 762–765, 2007.

[28] Joseph M Anand, S Guruprasad, and B Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *INTERSPEECH*, Pittsburgh, PA, USA, 2006, pp. 1–5.

[29] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[30] Thomas F Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education, 1^{st} (Eds.), 2006.

[31] Bayya Yegnanarayana and Hema A Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. on signal proc.*, vol. 40, no. 9, pp. 2281–2289, 1992.

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, United States, 2017, pp. 1125–1134.

[33] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016, pp. 1874–1883.

[34] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 6820–6824.

[35] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, 2017, pp. 1251–1258.

[36] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, The MIT Press, 1^{st} Eds., 2016.

[37] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, 5^{th} *Edition*, Information science and statistics. Springer, 2007.