INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Fake Audio Detection in Resource-constrained Settings using Microfeatures

Hira Dhamyal, Ayesha Ali, Ihsan Ayyub Qazi, Agha Ali Raza

Lahore University of Management Sciences, Pakistan

hira.dhamyal@lums.edu.pk, ayeshaali@lums.edu.pk, ihsan.qazi@lums.edu.pk,
agha.ali.raza@lums.edu.pk

Abstract

Fake audio generation has undergone remarkable improvement with the advancement in deep neural network models. This has made it increasingly important to develop lightweight yet robust mechanisms for detecting fake audios, especially for resource-constrained settings such as on edge devices and embedded controllers as well as with low-resource languages. In this paper, we analyze two *microfeatures*: Voicing Onset Time (VOT) and coarticulation, to classify bonafide and synthesized audios. Using the ASVSpoof2019 LA dataset, we find that on average, VOT is higher in synthesized speech compared to bonafide speech and exhibits higher variance for multiple occurrences of the same stop consonants. Further, we observe that vowels in CVC form in bonafide speech have greater F1/F2 movement compared to similarly constrained vowels in synthesized speech. We also analyse the predictive power of VOT and coarticulation for detecting bonafide and synthesized speech and achieve equal error rates of 25.2% using VOT, 39.3% using coarticulation, and 23.5% using a fusion of both models. This is the first study analysing VOT and coarticulation as features for fake audio detection. We suggest these microfeatures as standalone features for speaker-dependent forensics, voice-biometrics, and for rapid pre-screening of suspicious audios, and as additional features in bigger feature sets for computationally intensive classifiers.

Index Terms: Microfeature, voicing onset time, VOT, coarticulation, spoof, fake, bonafide, low-resource.

1. Introduction

Recent advances in deep learning have led to dramatic improvements in the automatic generation of natural sounding audios (e.g., using generative adversarial networks [1] and WaveNet based models [2]). This has led to their use in a wide variety of applications including in the design of assistive technologies, educational technologies, and games. Unfortunately, they can also be used as a powerful tool for spreading misinformation and for defeating automatic speaker verification (ASV) and voice-biometric systems. For example, audio deepfakes spreading false political narratives were regarded as a significant threat to the 2020 US presidential election [3]. Audio deepfakes have also been successfully used to fool ASV systems, where in one case it allegedly led to a loss of USD $243,000$ through a fraudulent bank transfer [3]. These examples highlight the extent of potential harm of natural sounding spoof audios as well as the vulnerability presented by the ASV systems, which are commonly used in many security systems.

To automatically distinguish between synthetically generated and bonafide audios, several data-driven and knowledge-driven countermeasure models have been proposed. While deep learning based models [4, 5, 6, 7] achieve high performance by automatically extracting discriminative features from audios, they require large amount of training data and computational resources. This makes such models a poor fit for resource-constrained settings like edge devices and IoT systems, where these models need to run on backend servers hence requiring internet connectivity and adding latency to the pipeline. Similarly, such models usually do not generalize well in sparse-data scenarios involving under-resource languages and forensic applications where data is harder to gather. On the other hand, knowledge-driven models exploit acoustic-level features such as the fundamental frequency, sequence-related entropy, and spectral envelope [8] as discriminative features.

Acoustic features may include *microfeatures*, which correspond to measurements that are made within the average duration of each unit of sound, typically within time intervals of the order of 1–20 ms [9]. These microfeatures are concomitant of the voice production mechanisms in humans which include the air pressure system, vibratory system, and the resonating system. The effects of all these systems exist at fine levels in the time and frequency domains. Bonafide audios carry such details, which we hypothesize are not replicated well in synthesized audios as such fine replication usually does not result in greater perceptual improvement. Microfeatures usually do not have direct perceptual correlates, i.e, they are difficult to capture by the untrained human ear and are also not captured by typical metrics used for evaluating synthesized audio.

In this work, we explore the effectiveness of two microfeatures, Voicing Onset Time (VOT) and coarticulation, for distinguishing between bonafide and spoofed speech. VOT is defined as the length of time that passes between the release of a stop consonant and the onset of voicing, the vibration of the vocal folds [10]. VOT has been shown to be correlated with age [11], speaking rate [12], diseases like Parkinson's [13], and even depression [14]. Due to differences across individuals, VOT can be utilized as a microfeature for speaker-dependent classification of audio files into bonafide and spoof.

Coarticulation refers to the influences of phonetic segments on adjacent or near-adjacent segments that are observable in the acoustic or articulatory patterns of speech [15]. In continuous speech, the anticipatory and residual movements of the articulators between adjacent phoneme configurations results in modified acoustic cues. The complex interplay of the articulators and the limited agility of the vocal tract to transform its shape from one configuration to another results in the mutual impact of surrounding phonetic units. This is manifested in the form of altered formant positions for vowels, and the assimilation of adjacent sounds to create altered places and manners of sound units. Coarticulation has been studied with respect to the different accents languages [16], how the Consonant-Vowel-Consonant (CVC) formant space for vowels changes in different languages [17], for better speech synthesis models [18] and in speech apraxia [19].

We hypothesize that spoof audios may not be able to accurately capture the VOT and coarticulation and thus by analysing

http://dx.doi.org/10.21437/Interspeech.2021-524

these features on a per-speaker basis, we can distinguish between bonafide and synthesized audio. We carry out our analysis using the ASVSpoof2019 Logical Access (LA) dataset [20], which consists of 6 different models for the generation of synthesized speech and 20 speakers. We find that the Random Forest (RF) based classifier achieves EER of 25.2% when VOT is used as the sole feature, 39.3% EER using coarticulation, and 23.5% with the fusion of scores from both models. Moreover, when VOT and coarticulation are used as input features in a deep learning model, they reduce the EER by 2%.

While deep neural networks can achieve much lower EERs compared to just using VOT and coarticulation based RF classifiers, the amount of data required for training may not be available (e.g., in case of low-resource languages) and their computational complexity makes it difficult to run them on edge devices (e.g., Amazon Echo). Therefore, we envision two use cases for our work: (a) the use of our microfeature based classifier as a lightweight filter that runs on low resource edge devices to flag suspicious audios (the audios can later be sent to the cloud for deeper analysis on larger neural network based models) and (b) as new input features to deep learning models for further improving their robustness.

2. Background Work

We now discuss two categories of related works. One category focuses on individual features whereas the other focuses on models for distinguishing between bonafide and spoof speech.

Gao et al., [8] study human voiced production based features including prosody features like jitter, shimmer and spectral entropy of F0 and found these features to be different between bonafide and spoof audio. Paul et al., [21] study short-term spectral features and found that lower frequencies ($<$ 1kHz) and high frequencies ($>$ 7kHz) are the most useful frequencies for discriminating between spoof and bonafide speech. Xiao et al., [22] use high dimensional magnitude and phase based features, and long term temporal information up to 0.51s. Relative phase information based features were exploited in [23], which out performed MFCC and MGDCC (modified group delay cepstral coefficients) on the dataset used. Short-term and long-term temporal modulation features were used in [24], which also performed better than MFCC and MGDCC. Leon et al., [25] use mean pitch stability, mean pitch stability range, and jitter as features extracted from image analysis of pitch patterns for discrimination between human and spoof speech using a classifier based on the Gaussian distributions of features. Todisco et al., [26] propose new features called CQCC (Constant Q cepstral coefficient) based on the constant Q transform. It provides a variable-resolution, time-frequency representation of the spectrum. Many deep learning countermeasure models are based on the CQCC features.

In terms of the models, Chintha et al., [4] use convolutional layers to extract features from the input audio. They use a wide block CNN architecture that consist of different kernel sizes to capture different levels of temporal dependencies. Chen et al., [5] use ResNet blocks for the input log filter banks of the audio, with data augmentation like reverberation and background noise. Halpern et al., [27] propose a CQT-ResNet and GMM architecture for bonafide and synthesized detection. Tak et al., [6] use an ensemble of classifiers that are trained on features of different spectral resolution in order to detect irregularities at either high or low frequencies bands. Hu et al., [28] propose squeeze and excitation, a CNN-based model, which has been shown to work well on this task.

In contrast to earlier works, we are the first to explore VOT and coarticulation as microfeatures for spoof audio detection. Such easy-to-compute features and the models based on them are useful in low resource conditions, where they can help prescreen suspicious audios. Such suspected audios could then be passed through more resource-heavy models. The proposed models can also be used for speaker-dependent forensics as well as to augment existing feature sets used in more computationally expensive spoof detection systems.

3. Microfeatures

3.1. Voicing Onset Time

VOT is of three types: (1) *0 VOT*, where the burst and voicing onset are spontaneous and near simultaneous, (2) *positive VOT*, where there is delay between burst and voicing onset, and (3) *negative VOT*, where voicing begins during the closure. VOT is a property of the stop consonants. There are 6 stop consonants in English, 3 voiced: /b d g/, and 3 voiceless: /p t k/. In some languages including English, stop consonants only obtain positive VOT of two types: a relatively long, positive VOT for voiceless stops, and a relatively short VOT for voiced stops [29]. Short-lag voiced stops have VOT ranging from 0 to $+25$ ms, with a median value of $+10$ msec. Long-lag voiceless stops have VOT ranging from $+60$ to $+100$ ms, with a median value of $+75$ ms [30]. We consider various factors when performing per-speaker analysis, e.g., the place of articulation of the stop phoneme, the following vowel, the following vowel height. Other factors including gender, age are not explored since this information is not present in the our dataset. Automatic measurement of VOT can be challenging because of the short duration of VOT, which is often between 10–20 ms. We measure the VOTs by using the AutoVOT software [31], which is a structured prediction algorithm described in [32].

3.2. Coarticulaiton

Coarticulation in an utterance can be captured by analysing the *formant* dynamics in a phoneme, where the formants represent the resonances of the vocal tract. Generally, formant changes are only analysed for *vowels* to capture coarticulation. In prior works, formant dynamics have been used for profiling; for example Loakes et al., [33] conclude that F3 is the most speaker-specific formant frequency range and for specific vowels it can be used to distinguish between similar sounding speakers.

To capture formant dynamics, we measure F1 and F2 for the duration of the vowel, and use this to calculate further measures of formant movements, including Vector Length (VL), Trajectory Length (TL), Trajectory Change (TC) and spectral Rate of Change (roc). VL is the length of the vector in the F1/F2 plane and is an indication of the amount of formant change over the course of the vowel. Given that formants are calculated at n points over the duration of the vowel, VL is calculated as $\mathrm{VL} = \sqrt{(\mathrm{F1}_1 - \mathrm{F1}_n)^2 + (\mathrm{F2}_1 - \mathrm{F2}_n)^2}$. TL takes a closer look into the change of the formants than VL and is calculated as: $\mathrm{TL} = \sum_{m=1}^{n} \mathrm{VSL}_m$, where $\mathrm{VSL}_m = \sqrt{(\mathrm{F1}_m - \mathrm{F1}_{m+1})^2 + (\mathrm{F2}_m - \mathrm{F2}_{m+1})^2}$. Formant TC is calculated as: $\mathrm{TC} = \sum_{m=1}^{n} (\mathrm{F1}_m - \mathrm{F1}_{m+1}) + (\mathrm{F2}_m - \mathrm{F2}_{m+1})$. Spectral roc is defined as the change of the TL over the duration of the vowel: $\mathrm{TL\ roc} = \frac{\mathrm{TL}}{\mathrm{duration}}$. We calculate and compare these metrics from the spoof and bonafide audios. For each vowel, we take into consideration one left and right neighbouring consonant. Since coarticulation is affected by several factors, including the prosodic overlays [34], language [17], context of the phoneme, and speaker; we do a per-speaker analysis only in

Table 1: *Spoof audio generation systems in ASVSpoof2019 LA. [35]*

System	Description
A01	NN based TTS system using the VAE-LSTM as acoustic model and Wavenet vocoder for waveform generator
A02	NN based TTS system using the VAE-LSTM as acoustic model and WORLD vocoder for waveform generator
A03	Feedforward NN as acoustic model and WORLD vocoder for waveform generator
A04	A waveform concatenation TTS
A05	NN VC system using VAE as VC model and WORLD vocoder for waveform generator
A06	Transfer function based VC system
Total	No. of speakers: 20, bonafide: 2580, spoof: 22800

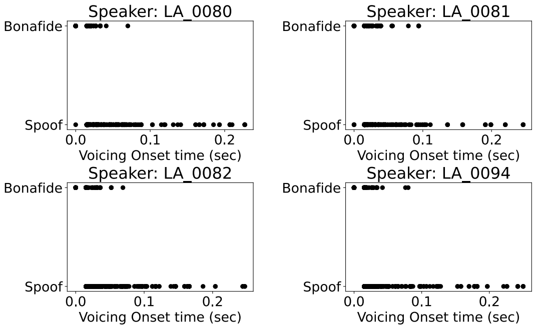

Figure 1: *Scatter plot of the VOT for spoof and bonafide audios for four speakers in the ASVspoof2019 LA dataset. All of the stop phonemes are included in this scatter plot.*

English taking into consideration the neighbouring consonants.

4. Experimental Evaluation

4.1. Dataset

To analyse the VOT and coarticulation, we use the training section of the ASVSpoof2019 LA dataset [20]. The spoof audio samples in this dataset are created using a set of text to speech (TTS) systems and voice conversation (VC) algorithms detailed in Table 1.

4.2. Methodology

The AutoVOT library requires word aligned transcript and the audio signal as input. Therefore, we first use Google ASR API [36] to transcribe the audio and then use an HMM-based phoneme segmentor [37] to get word and phone boundaries. Each audio sample and each word boundary are separately passed through the AutoVOT model to get the boundaries of predicted VOT. For formant calculation, we use the formant tracking algorithm in praat [38], which uses a 25 ms analysis window (Gaussian), and computes LPC coefficients.

4.3. VOT Analysis

Fig. 1 shows the scatter plot of VOT and the label, spoof or bonafide, for four speakers in the dataset, with VOT on the x-axis and label on the y-axis. Observe that while there is an overlap in the spoof and bonafide VOTs, spoof audios exhibit much larger variations than bonafide audios. In particular, the VOT of the stop phonemes ranges from 0 to 100 ms in case of bonafide audio, whereas for spoof audios, it ranges from 0 to 250 ms. Similar trend was observed for all speakers.

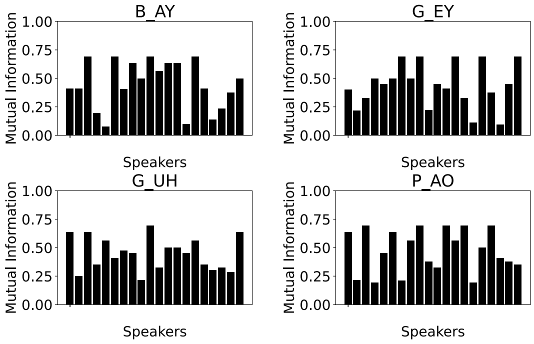

Figure 2: *Mutual information score between the VOT and label spoof and bonafide.*

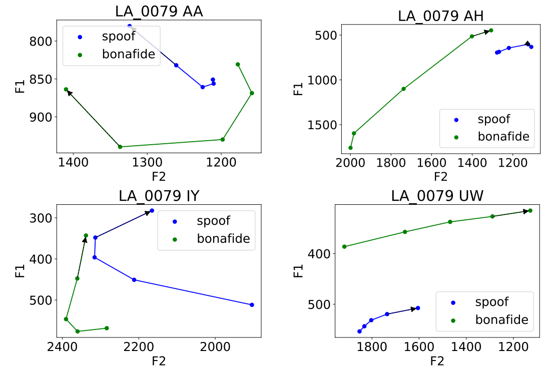

Figure 3: *Formant movements for 4 vowels for a single speaker.*

Fig. 2 shows the mutual information (MI) between VOT and the labels for four stop consonants and vowel combination. MI for a particular phoneme like /g/ followed by vowel /uh/ is not consistent over different speakers. For example, for speaker LA-0088, MI score is 0.7 and for speaker LA-0087, it is 0.2. No single phoneme alone has good MI over all the speakers.

We also performed a per spoof generation system analysis, where for each speaker, we divide the audios further into per system category. The macro F1 of a RBF when trained using all stop phonemes ranges from 58.9% to 64.0% - averaged across all speakers. We observe that the vot model performs particularly well for system A05 across all speakers (64.0%) ; on which the performance increases to 68.3% when only considering the phonemes where following vowel height is low.

4.4. Coarticulation Analysis

For the bonafide and spoof utterances, we compare the actual F1/F2 values, average F1/F2 difference in successive frames, velocity and acceleration in an utterance, VL, TL and TL roc. All the above metrics measure the *change* of the formants over the duration of the vowel. We find that on average, the above metrics are higher for bonafide than for spoof. Performing unpaired t test on the metrics gives the following results: F1 velocity: p-value $= 0.00e{-}06$, F2 velocity: p-value $= 0.00e{-}06$, F1 acceleration: p-value $= 0.00e{-}04$, F2 acceleration: p-value $= 0.00e{-}04$, VL: p-value $= 0.00e{-}16$, TL: p-value $= 0.00e{-}13$, TL-roc: p-value $= 0.00e{-}15$ at alpha $= 0.1$. All of the above metrics of coarticulation show a statistically significant result. Fig. 3 shows some of the differences in the formant movement for selected phonemes in bonafide and spoof utterances. The figure illustrates higher rate of formant movements in bonafide as compared to the spoof utterance. A per system analysis using F1 velocity results in F1 scores of a RBF model ranging from 61.9% to 63.7%.

4.5. Classification results

We demonstrate here the predictive power of using VOT alone as a feature to distinguish between spoof and bonafide. Each word consisting of a stop phoneme and a detected VOT is considered as a separate instance. To use VOT as a feature, we make one hot vectors the size of the total number of different phonemes that are being used. For each instance of stop and vowel combination, we add the VOT value to the one hot vector at the index of that stop-vowel combination. Using this feature, we train several machine learning models separately for each of the 20 speakers. We perform 5-fold cross validation and get the average result over the folds for each speaker. Table 2 shows the F1 macro averaged over all the 20 speakers, when using different subsets of stop phoneme-vowel combination for 3 models.

The table shows the highest F1-macro achieved using RF on the subset *E* which consists of all the stops phonemes followed by a high vowel. It is expected that VOT for stop phonemes following a high vowel is longer than following a non-high vowel. Since the average VOT for spoof audio is higher than the average VOT for bonafide audio, phone-vowel combination with shorter VOT are deemed better distinguishing factor among bonafide and spoof. This also explains how VOT for the following vowel height=low performs better than when the following vowel height is mid and high. This suggests that VOT alone can be a distinguishing feature.

Table 2: *Macro F1 achieved using RBF, SVM and KNN classifiers over different subsets of the features used. Subset A consists of* Voiced stops followed by any vowel. *B consists of* Voiceless stops followed by any vowel. *C consists of* all stops followed by vowel height=high. *D consists of* all stops followed by vowel height=mid. *E consists of* all stops followed by vowel height=low. *F consists of* stops with manner of articulation=bilabial. *RF gives the best results.*

Model	Macro F1 on different subsets of phonemes					
	A	*B*	*C*	*D*	*E*	*F*
Random Forest	67.2	69.1	64.8	67.7	**71.5**	69.6
SVM	58.5	57.3	53.4	57.3	64.8	62.8
KNN	58.9	63.3	52.2	59.2	53.8	57.5

We also test the predictive power of the coarticulation metrics described before. Similar to VOT, we use one hot vectors to encode the features for each CVC phoneme combinations and train random forest, SVM and KNN models on each of the 20 speakers and perform cross validation. Table 3 shows that the F1 velocity over the duration of the vowel is the most distinguishing feature. Spoof audio lacks the smooth transitions of the phonemes, which is also audible from the utterances and we expect the velocity and acceleration values to capture such distinctions between the spoof and bonafide utterances.

Finally, we compare the benefit of using VOT and the different measures of coarticulation in a Convolutional Neural Network model. The model consists of 5 CNN layers, with [1, 4, 16, 64, 256] feature channels in each layer, trained using binary cross entropy loss. (To be able to use VOT and coarticulation features, we only consider the audio files from which we could gather both of these features.) Table 4 shows the results. We first train the model using CQCC features alone and achieve 5% EER. When using VOT and coarticulation features together with CQCC, we achieve an EER of 3%. Also observe that RF based models using VOT, Coarticulation, and combined features only require 0.1K-2K FLOPS (Floating Point Operations per sec) compared to 32M FLOPS with the CNN models. This makes the former particularly suitable to run on resource-constrained edge devices.

Table 3: *Macro F1 achieved using RBF, SVM and KNN classifiers using different metrics of coarticulation.*

Model	Macro F1		
	f1 velocity	*f2 velocty*	*TC*
Random Forest	**67.7**	67.0	66.7
SVM	51.9	53.3	53.9
KNN	51.9	53.3	50.1

Table 4: *EER using various models.*

Model	RF	RF	RF-Fusion	CNN	CNN
Features	Coart	VOT	Both	CQCC	CQCC + coart + vot
EER(%)	39.3	25.2	23.5	5.0	3.0
FLOPs	1.5K	0.1K	2K	32M	32M
Params	-	-	-	0.2M	0.2M

5. Limitations and Conclusion

There are two sources of variability in the pipeline we use in our work. First, using a pronunciation dictionary to represent each word into a series of phonemes, may sometimes create a difference in the expected phoneme occurrence in the audio sample and the actual phonemes present in the pronunciation, which can affect the detection of VOT and formant analysis. However, given the large number of total utterances used in our analysis, we expect such discrepancies to be rare and unlikely to significantly affect the classification results. Second, the AutoVOT system used to detect VOT also has an average error rate of 5 msec. However, this is much smaller than the range of VOTs exhibited by both spoof and bonafide audios.

From our experiments, we conclude that VOT and coarticulation are discriminative microfeatures, more so VOT. Spoof speech can exceed the normal range of VOT and on average is higher than the VOT in bonafide speech. The abnormality of VOT in spoof speech is an indicator of the fine details in speech that synthesized speech systems have not been able to capture so well. Furthermore, formant movements for vowels, occurring in between consonants (CVC), is lower in spoof than in bonafide. We have developed simple models on easily measurable microfeatures from voice, which may act as a low resource model for audio spoof detection. Our results show that such models could act as relatively cheaper, quicker and reliable screening mechanism before bringing in the cloud-based deep learning models, consisting of millions of parameters. The systems are advancing in creating more natural sounding speech by capturing the macro details of speech, which may be able to fool the human ear. However, our work shows that inconsistencies of spoof audio also lie in the finer details of speech, that the human ear is not trained to capture. Spoof audio generation is continually advancing but replicating the fine level details of the human voice production mechanism is still far from reach, and not even an identified goal for now, and this paper is an effort in exploring two such micro-level features.

6. References

[1] Y. Gao, R. Singh, and B. Raj, “Voice impersonation using generative adversarial networks,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2506–2510.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[3] K. Hartmann and K. Giles, “The next generation of cyber-enabled information warfare,” in *2020 12th International Conference on Cyber Conflict (CyCon)*, vol. 1300. IEEE, 2020, pp. 233–250.

[4] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, “Recurrent convolutional structures for audio spoof and video deepfake detection,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.

[5] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, “Generalization of audio deepfake detection,” in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 132–137.

[6] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, “Spoofing attack detection using the non-linear fusion of sub-band classifiers,” *arXiv preprint arXiv:2005.10393*, 2020.

[7] Z. Wu, R. K. Das, J. Yang, and H. Li, “Light convolutional neural network with feature genuinization for detection of synthetic speech attacks,” *arXiv preprint arXiv:2009.09637*, 2020.

[8] Y. Gao, J. Lian, B. Raj, and R. Singh, “Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems,” *arXiv preprint arXiv:2011.03689*, 2020.

[9] R. Singh, *Profiling humans from their voice*. Springer, 2019.

[10] L. Lisker and A. S. Abramson, “The voicing dimension: Some experiments in comparative phonetics,” in *Proceedings of the 6th international congress of phonetic sciences*, vol. 563. Academia Prague Prague, Czech Republic, 1970, pp. 563–567.

[11] R. J. Morris and W. Brown Jr, “Age-related differences in speech variability among women,” *Journal of Communication Disorders*, vol. 27, no. 1, pp. 49–64, 1994.

[12] K. Stölten, N. Abrahamsson, and K. Hyltenstam, “Effects of age and speaking rate on voice onset time,” *Studies in Second Language Acquisition*, vol. 37, no. 1, p. 71, 2015.

[13] E. Fischer and A. M. Goberman, “Voice onset time in parkinson disease,” *Journal of Communication Disorders*, vol. 43, no. 1, pp. 21–34, 2010.

[14] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, “Acoustic analysis in the differentiation of parkinson’s disease and major depression,” *Journal of Psycholinguistic Research*, vol. 21, no. 5, pp. 383–399, 1992.

[15] C. A. Fowler and E. Saltzman, “Coordination and coarticulation in speech production,” *Language and speech*, vol. 36, no. 2-3, pp. 171–195, 1993.

[16] G. Zellou, R. Scarborough, and R. Kemp, “Secondary phonetic cues in the production of the nasal short-a system in california english,” *Proc. Interspeech 2020*, pp. 631–635, 2020.

[17] P. S. Beddor, J. D. Harnsberger, and S. Lindemann, “Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates,” *Journal of Phonetics*, vol. 30, no. 4, pp. 591–627, 2002.

[18] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” *PloS one*, vol. 8, no. 4, p. e60603, 2013.

[19] W. Ziegler and D. Von Cramon, “Anticipatory coarticulation in a patient with apraxia of speech,” *Brain and Language*, vol. 26, no. 1, pp. 117–130, 1985.

[20] “Asvspoof 2019: the automatic speaker verification spoofing and countermeasures challenge evaluation plan.” http://www.asvspoof.org/asvspoof2019/asvspoof2019 evaluation plan.pdf, 2019, [Online].

[21] D. Paul, M. Pal, and G. Saha, “Spectral features for synthetic speech detection,” *IEEE journal of selected topics in signal processing*, vol. 11, no. 4, pp. 605–617, 2017.

[22] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, “Spoofing speech detection using high dimensional magnitude and phase features: The ntu approach for asvspoof 2015 challenge,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[23] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, “Relative phase information for detecting human speech and spoofed speech,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] Z. Wu, X. Xiao, E. S. Chng, and H. Li, “Synthetic speech detection using temporal modulation feature,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7234–7238.

[25] P. L. D. Leon, B. Stewart, and J. Yamagishi, “Synthetic speech discrimination using pitch pattern statistics derived from image analysis,” in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[26] M. Todisco, H. Delgado, and N. W. Evans, “A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients.” in *Odyssey*, vol. 2016, 2016, pp. 283–290.

[27] B. M. Halpern, F. Kelly, R. van Son12, and A. Alexander, “Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure.”

[28] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[29] M. Deuchar and A. Clark, “Early bilingual acquisition of the voicing contrast in english and spanish,” *Journal of Phonetics*, vol. 24, no. 3, pp. 351–365, 1996.

[30] P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, “Voice onset time in aphasia, apraxia of speech and dysarthria: a review,” *Clinical linguistics & phonetics*, vol. 14, no. 2, pp. 131–150, 2000.

[31] J. Keshet, M. Sonderegger, and T. Knowles, “Autovot: A tool for automatic measurement of voice onset time using discriminative structured prediction [computer program],” *Version 0.91, retrieved November*, 2014.

[32] M. Sonderegger and J. Keshet, “Automatic measurement of voice onset time using discriminative structured prediction,” *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3965–3979, 2012.

[33] D. Loakes, “A forensic phonetic investigation into the speech patterns of identical and non-identical twins,” *International Journal of Speech Language and the Law*, vol. 15, no. 1, pp. 97–100, 2008.

[34] B. Lindblom, A. Agwuele, H. M. Sussman, and E. E. Cortes, “The effect of emphatic stress on consonant vowel coarticulation,” *The Journal of the Acoustical Society of America*, vol. 121, no. 6, pp. 3802–3813, 2007.

[35] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee *et al.*, “Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[36] “Google speech to text api,” https://cloud.google.com/speech-to-text/.

[37] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, “The cmu sphinx-4 speech recognition system,” in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, vol. 1, 2003, pp. 2–5.

[38] “Praat.” https://www.fon.hum.uva.nl/praat, [Online].