Speaker Verification Enhancement via Speaking Rate Dynamics in Persian Speechprints

Nina Hosseini-Kivanani[1][a], Homa Asadi[2][b] and Christoph Schommer[1][c]

[1]*Department of Computer Science, University of Luxembourg, Esch-sur-Alzette, Luxembourg*

[2]*Faculty of Foreign Languages, University of Isfahan, Isfahan, Iran*

Keywords: Speaker Verification, Mel-Frequency Cepstral Coefficients (MFCCs), Vowel Formants, Deep Learning, Persian Language.

Abstract: This paper investigates the impact of speaking rate variation on speaker verification using a hybrid feature approach that combines Mel-Frequency Cepstral Coefficients (MFCCs), their dynamic derivatives (delta and delta-delta), and vowel formants. To enhance system robustness, we also applied data augmentation techniques such as time-stretching, pitch-shifting, and noise addition. The dataset comprises recordings of Persian speakers at three distinct speaking rates: slow, normal, and fast. Our results show that the combined model integrating MFCCs, delta-delta features, and formant frequencies significantly outperforms individual feature sets, achieving an accuracy of 75% with augmentation, compared to 70% without augmentation. This highlights the benefit of leveraging both spectral and temporal features for speaker verification under varying speaking conditions. Furthermore, data augmentation improved the generalization of all models, particularly for the combined feature set, where precision, recall, and F1-score metrics showed substantial gains. These findings underscore the importance of feature fusion and augmentation in developing robust speaker verification systems. Our study contributes to advancing speaker identification methodologies, particularly in real-world applications where variability in speaking rate and environmental conditions presents a challenge.

1 INTRODUCTION

Speech production is a highly complex phenomenon in which dynamic articulatory gestures drive the movements of speech organs to achieve specific targets within the vocal tract geometry (Tilsen, 2014). These articulatory movements shape the acoustic features of speech, which carry rich information, enabling listeners to comprehend both the linguistic content (what is said) and the speaker-specific details (who said it). Identifying speakers based on the characteristics of their voices is a fundamental goal in forensic phonetics and automatic speaker recognition systems (Rose, 2002; Nolan, 1987).

One of the key challenges in speaker identification is the high variability in acoustic characteristics across speakers, compared to the relatively low variability within a single speaker (Gold et al., 2013; McDougall, 2006). In forensic speaker comparison

(FSC), addressing this variability is crucial when determining whether a known voice sample matches an unknown (disputed) sample (Rose, 2002; Nolan, 1987).

Identifying robust speaker-specific parameters remains challenging due to intertwined factors, such as linguistic effects, prosody, and channel conditions, which influence system accuracy (Rose, 2002). Among these factors, speaking rate variability plays a particularly important role, as speakers naturally adjust their rate based on communicative context, emotional state, or physiological conditions (Gay et al., 1974; Imaizumi and Kiritani, 1989). Such variations affect both articulatory movements and acoustic properties, posing challenges to speaker recognition systems (Shahrebabaki et al., 2018; Zeng et al., 2015).

From a computational standpoint, automatic speaker identification (ASI) systems must account for changes in speaking rate (Reynolds et al., 2000). Traditional ASI systems use MFCCs as their primary feature set due to their effectiveness in capturing speaker-specific spectral information (Davis and Mermelstein, 1980). However, these systems often degrade when

faced with speaking rate variations, as spectral characteristics shift with changes in articulatory dynamics(Zeng et al., 2015). Recent advances in deep learning (DL) have addressed this by developing speaker-invariant representations that are more robust to such variations (Hinton et al., 2012; Xie et al., 2019).

Despite these advancements, vowel formants remain valuable in forensic applications due to their interpretability and close relationship to the physiological aspects of speech production (Gold et al., 2013). Given that both vowel formants and MFCCs are differentially affected by speaking rate, assessing their relative robustness is critical for improving speaker identification systems.

In this study, we evaluate the impact of speaking rate variability on speaker identification performance by comparing the robustness of vowel formants and MFCCs. Specifically, we aim to:

1. examine how speaking rate influences vowel formant frequencies and MFCC features;

2. assess the effectiveness of these features in speaker identification under varying speaking rates;

3. and determine whether combining formant and MFCC features improves accuracy and resilience in speaker verification systems.

By addressing these objectives, this study provides insights into selecting acoustic features most effective for speaker identification under conditions of variable speaking rates. The remainder of this paper is organized as follows: Section 2 reviews related work from both phonetic and computational perspectives. Section 3 outlines our experimental methodology. Section 4 presents our findings, while Section 5 discusses their implications. Finally, Section 6 concludes the paper and suggests directions for future research.

2 RELATED WORK

Forensic speaker comparison (FSC) and automatic speaker recognition systems rely on acoustic features, such as vowel formants and Mel-Frequency Cepstral Coefficients (MFCCs), to differentiate speakers based on their unique vocal characteristics (Rose, 2002; Nolan, 1987). These features effectively capture speaker-specific and linguistically relevant aspects of speech, making them central to forensic and automatic speaker verification tasks.

Speaking rate is a critical variable that affects the articulation and acoustic properties of speech. Research has shown that varying speaking rates influ-

ence the kinematics of speech production. For instance, Gay (Gay et al., 1974) demonstrated that increased speaking rate is associated with heightened muscle activity, such as more pronounced lip closure and greater bilabial consonant openings. Similarly, Tuller and Kelso (Tuller et al., 1982) found that faster-speaking rates result in shorter muscle activity durations, while slower rates lead to longer articulatory movement times. These findings underscore how speaking rate directly impacts speech articulation, making a vital factor in speaker verification tasks. These articulatory changes, influenced by speaking rate, are further reflected in the variability of specific speech gestures. For instance, Shaiman et al. (Shaiman et al., 1997) observed that lip gestures exhibit variability across different speaking rates, complicating the consistency of speaker-specific features as articulation velocity changes.

From an acoustic perspective, speaking rate significantly alters the spectral characteristics and formant trajectories of speech. Imaizumi and Kiritani (Imaizumi and Kiritani, 1989) observed that rapid speech can lead to vowel reduction, especially in the second formant frequency (F2). Additionally, Weismer and Berry (Weismer and Berry, 2003) showed that speakers modify formant movement trajectories based on their speaking rate, with F2 being particularly affected by these changes. This suggests that variations in speaking rate not only affect articulatory gestures but also modify key acoustic features critical for speaker recognition.

Furthermore, research by Mefferd and Green (Mefferd and Green, 2010) demonstrated that formant transitions become sharper at higher speaking rates. Their studies also noted that vowel formant distances exhibit greater specificity in slow speech compared to fast speech. Agwuele et al. (Agwuele et al., 2009) found a reduction in vowel space with faster speaking rates, indicating that articulation and vowel acoustics are closely tied to the speed of speech. Together, these studies highlight the variability in formant frequency patterns and their dependency on speaking rate, a critical challenge for speaker identification systems.

Recent advancements in speaker verification have explored the use of learnable MFCCs to improve robustness to variable speaking rates. For instance, Liu et al. (Liu et al., 2021) introduced adaptive MFCC front-end architectures that adjust to data, making these features more resilient to changing speech conditions, including speaking rate variability. This adaptive approach has shown significant improvements in speaker verification performance, particularly in large-scale datasets like VoxCeleb1 and SITW. How-

ever, while learnable MFCCs offer improvements, they may not fully capture the dynamic properties of speech under all conditions.

2.1 Speaker Recognition and Computational Approaches

In traditional automatic speaker recognition, MFCCs are widely used for their robustness in capturing speaker-specific spectral properties (Davis and Mermelstein, 1980; Reynolds et al., 2000). However, they are susceptible to degradation under varying speaking rates, as spectral characteristics shift due to articulatory dynamics (Zeng et al., 2015). Zeng and Sheng (Zeng et al., 2015) demonstrated that these changes directly affect the reliability of MFCC-based systems. To address these limitations, recent approaches have incorporated machine learning (ML) and deep learning (DL) techniques, such as convolutional neural networks (CNNs) and utterance-level aggregation methods (Xie et al., 2019), which learn hierarchical representations that improve the robustness of speaker verification systems under speaking rate variability (Hinton et al., 2012).

Recent work has also explored hybrid models that integrate the features of MFCCs with DL architectures. For instance, combining MFCCs with CNNs captures both local and global speech patterns, enhancing robustness to noise and rate variability. Furthermore, bi-directional long short-term memory (Bi-LSTM) networks improve speaker verification by modeling long-range temporal dependencies, which are essential for capturing dynamic speech variations (Anupama et al., 2022).

Hybrid optimization strategies have further advanced these models. Chakravarty & Dua (Chakravarty and Dua, 2023) combined MFCCs and Gammatone Cepstral Coefficients (GTCCs) with data augmentation methods, such as Synthetic Minority Over-Sampling Technique (SMOTE), to improve performance. By leveraging a hybrid LSTM backend, their approach enhanced model accuracy and resilience under noisy conditions, demonstrating the value of combining feature sets and optimization algorithms in varying conditions.

In forensic applications, interpretability remains critical, making vowel formants valuable despite their susceptibility to speaking rate variability. Formants offer insights into the physiological aspects of speech production, which are useful for distinguishing speakers (Asadi et al., 2018; McDougall, 2006). However, their limitations as standalone features necessitate combining formants with robust features like MFCCs. Jahangir et al. (Jahangir et al., 2020)

demonstrated that integrating traditional acoustic features with DL-derived representations significantly enhances speaker identification under variable speaking rates.

Some studies have also shown that the fusion of acoustic features improves performance. Bahari et al.(Bahari and Van Hamme, 2011) demonstrated that combining formant frequencies and MFCCs captures both articulatory and spectral information, leading to better recognition rates under challenging conditions. Advanced modeling techniques, such as i-vectors and x-vectors, have further demonstrated robustness across varying speaking conditions(Dehak et al., 2010). These findings emphasize the benefits of combining diverse feature sets for robust speaker verification systems.

2.2 Research Gap

While progress has been made in understanding how speaking rate affects speech production and acoustic features, few studies comprehensively compare the robustness of formants and MFCCs in speaker verification under varying rates. Further research is needed to evaluate whether combining these features with deep learning techniques can improve resilience to speaking rate variability, especially in forensic and real-world applications. In this study, we address this gap by systematically examining the impact of speaking rate on formant frequencies and MFCCs in speaker verification tasks. Additionally, we explore the benefits of integrating these features into a unified framework to improve accuracy and robustness under variable speaking conditions.

3 EXPERIMENTAL SET-UP

3.1 Participants and Task

Eighteen male Persian speakers (Tehrani variety; age range: 25–36 years; M = 31.3, SD = 3.7) were recorded. None of the speakers reported any hearing or speech impairments. All participants were students pursuing a master's or PhD degree in various research areas. This corpus was collected following the procedure used in the collection of the BonnTempo corpus in German. Speakers were instructed to read The North Wind and the Sun in Persian at three different speaking rates (slow, normal, and fast). Before each recording session, participants were asked to read the text several times to familiarize themselves with the passage. First, speakers were instructed to read the passage at their normal pace. The speakers were then

asked to slow their pace as much as they could and then to read the text as fast as possible. This resulted in strong syllable rate variability across the three different reading passages. All recording sessions took place in a soundproof booth with a sampling rate of 44.1 kHz and 16-bit quantization.

3.2 Feature Extraction

Speech recordings were labeled and segmented based on the onset and offset information using Praat (Boersma and Weenink, 2021) version 6.2.22. A free plugin for Praat with automated scripts for voice processing, Praat Vocal Toolkit (Corretge, 2022), was used to extract and concatenate all vowels from each recording per speaker. Formant values were extracted at 5-ms intervals using the LPC-based Burg algorithm in Praat. A long-term analysis method was adopted, as it has proven effective in representing speaker individuality (Asadi et al., 2018; Gold et al., 2013). This approach calculates the average formant values over a long stretch of a speaker's speech recording (Gold et al., 2013; Rose, 2002; Nolan, 1987).

The key features were extracted using MFCCs in Python (Version 3.11.5) with the librosa library (McFee et al., 2015). Thirteen main coefficients were calculated and averaged per audio file to simplify the representation while preserving key sound characteristics.

We developed a multi-class speech analysis model that extensively uses MFCCs, delta-MFCCs, and delta-delta-MFCCs to capture a broad spectrum of acoustic features from speech. This strategy enriches the model with spectral and temporal information, improving its ability to distinguish between varied speaking speeds and speaker characteristics. The methodology centers on extracting a rich set of features from audio recordings: the base MFCCs provide spectral information, delta-MFCCs capture the rate of change in these spectral features, and delta-delta-MFCCs further detail the acceleration of these changes. This layered approach to feature extraction ensures a deep representation of the audio's characteristics. Each feature dimension is normalized to ensure consistency in scale across the dataset, facilitating more effective model training.

Adding attention mechanisms, such as self-attention or Transformer layers, can help the model focus on the most relevant parts of the speech signal for speaker verification, making it more effective in handling variations in speaking speed. We then transformed these labels into a one-hot encoded format suitable for classification tasks. The dataset was split

into training and testing sets, with 20% reserved for testing to evaluate the model's performance. Our neural network model architecture included Long Short-Term Memory (LSTM) layers to process the sequential nature of the audio data, followed by a custom AttentionLayer designed to weigh the importance of different parts of the audio signal, enhancing the model's focus on relevant features for speaker verification. The model also incorporated dropout layers to prevent overfitting and used a softmax activation function in the output layer for classification. The model was compiled using the Adam optimizer and categorical cross-entropy loss. Early stopping was implemented to terminate training when validation loss ceased to improve, thereby preventing overfitting

We used Group K-Fold Cross-Validation, a variant of K-Fold, to divide the dataset into five folds while ensuring that all recordings from a single speaker were placed either in the training or testing set, not both. This approach prevents data leakage and ensures a fair evaluation by maintaining class proportions across folds, allowing the model to generalize effectively to unseen speakers.

3.3 Speech Data Augmentation

To enhance dataset diversity and improve model generalization, five audio augmentation techniques were applied using the librosa library. Augmentation introduces variability in training data, helping models better handle real-world challenges such as varying speaking rates, background noise, and recording conditions (Lounnas et al., 2022).

Time-Stretching: We adjusted the speed of the audio using factors of 0.9 (slower) and 1.1 (faster). This manipulation simulates natural variations in speaking rate, allowing the model to adapt to speakers with different articulation speeds (Ko et al., 2015).

Pitch Shifting: The pitch of the audio was shifted by ±2 semitones to mimic variations in vocal pitch, which may occur due to speaker differences or emotional states. This augmentation captures variations in speaker intonation while maintaining the original speech content (Alex et al., 2023).

Noise Addition: Gaussian noise was added to the audio with an amplitude factor of 0.005 to simulate environmental noise. This method increases robustness by enabling the model to process noisy input data, reflecting real-world recording conditions (Nugroho et al., 2021).

Volume Adjustment: The amplitude of the audio was scaled to 80% and 120% of its original level to simulate variations in the recording conditions and speaker distance from the microphone. These adjust-

ments ensure that the model can handle varying input signal strengths (Zhou et al., 2017).

Audio Shifting: A random circular shift of up to 20% of audio length was applied to simulate misalignments or varying speech onset times, enhancing the model's ability to handle such variances (Lounnas et al., 2022).

Each augmented audio file was saved as a separate sample, effectively increasing the dataset size and introducing acoustic and temporal variability. This augmentation strategy ensured that the model became more resilient to natural variations and real-world challenges. The augmented dataset was subsequently used for model training, and the impact of each augmentation type was evaluated for its contribution to system performance.

3.4 Model Training and Evaluation

To assess the performance of our models, we computed classification accuracy, precision, recall, and F1-score. These metrics are standard in classification tasks and provide insights into the model's overall correctness (accuracy), the proportion of correctly identified positive cases (precision), the ability to identify all relevant positive cases (recall), and a balanced measure of precision and recall (F1-score). Detailed definitions of these metrics are available in the standard machine learning literature.
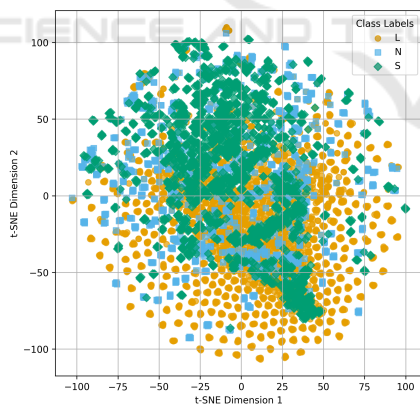

Figure 1: t-Sne visualization of speaker embeddings derived from MFCC, formant, and pitch features across three classes of speaking rates: L (Low), N (Normal), and S (Speedy).

Figure 1 presents a t−SNE projection of the speaker embeddings derived from multiple acoustic features, including MFCCs, formant frequencies, and pitch, across three distinct speaking rates. The visual separation of the clusters demonstrates the discriminative power of the combined feature set, particularly in capturing the temporal variations in speech

signals. Notably, the dense overlap between some points highlights the inherent challenges of speaker verification under varying speaking rates. This result reinforces the importance of integrating complementary features such as MFCCs and formants to enhance the robustness of speaker verification systems, especially in real-world applications where speaking rate variability is prevalent.

4 RESULTS AND DISCUSSION

We used 5-fold stratified cross-validation for model training, running each fold for 50 epochs with a learning rate of 0.0001 and a batch size of 32. Early stopping with a patience of 10 epochs was applied to prevent overfitting and to retain optimal model weights. The models were optimized using the Adam optimizer and categorical cross-entropy loss for multi-class classification tasks.

Table 1 presents the performance metrics—accuracy, precision, recall, and F1-score—of different acoustic feature sets: MFCC, MFCC-delta-delta, and formant frequencies (F0, F1–F4), evaluated under augmented and non-augmented conditions.

Without Augmentation: The combined model (MFCC, formant frequencies, and MFCC derivatives) achieved the highest performance across all metrics, with an accuracy of 70%, precision of 69%, recall of 67%, and F1-score of 68%. This result highlights the advantage of combining diverse acoustic features to capture speaker-specific information. In contrast, individual feature sets performed lower: MFCC-delta-delta achieved 62% accuracy, MFCC alone reached 60%, and formant frequencies performed the weakest at 59%, underscoring their limited discriminatory power when used in isolation.

With Augmentation: all models showed clear improvements, confirming its role in enhancing model robustness. The combined model again outperformed individual models, achieving 75% accuracy, with precision, recall, and F1-scores of 74%, 74%, and 73%, respectively. Augmentation also notably improved individual feature sets. Formant frequencies exhibited a significant increase in accuracy, rising to 68%, matching the MFCC-delta-delta model, which improved to 67% accuracy. Even MFCC alone benefited, achieving 65% accuracy compared to 60% without augmentation.

These results underscore the effectiveness of combining spectral (MFCC), temporal (delta and delta-delta), and source-related (formants) features to provide a comprehensive representation of speaker characteristics. The combined model consistently out-

Table 1: Model performance of MFCC, F0, and Formant Frequencies, and their combinations with and without augmentation.

Augmentation	Model	Accuracy	Precision	Recall	F1-score
Without Aug.	MFCC	60%	61%	60%	59%
	MFCC-delta-delta	62%	60%	63%	62%
	F0, F1-F4	59%	56%	55%	55%
	Combined Model	**70%**	**69%**	**67%**	**68%**
With Aug.	MFCC	65%	62%	62%	61%
	MFCC-delta-delta	67%	64%	63%	64%
	F0, F1-F4	68%	68%	67%	67%
	Combined Model	**75%**	**74%**	**74%**	**73%**

performed individual feature sets in both conditions, while data augmentation further enhanced model performance, introducing variability that simulates real-world speaking conditions. The superior results achieved by the augmented combined model demonstrate the robustness and generalizability of this multi-feature approach, particularly in handling variations in speaking styles and conditions, making it a strong foundation for speaker verification systems.

5 DISCUSSION

The findings from this study highlight the impact of combining multiple acoustic feature sets and applying data augmentation on the robustness and accuracy of speaker identification systems. Specifically, the combined model (MFCC, MFCC-delta, MFCC-delta-delta, and formant frequencies) outperformed individual feature models in both augmented and non-augmented conditions, achieving a 75% accuracy with augmentation, compared to 70% without augmentation. This confirms previous work suggesting that hybrid approaches, which leverage both spectral and temporal information, provide a more comprehensive representation of speaker-specific traits (Dehak et al., 2010).

One reason for the success of MFCC-based features in this context is their ability to capture speaker-specific spectral envelopes, which have long been considered the foundation for speaker recognition tasks (Davis and Mermelstein, 1980). The delta and delta-delta coefficients add temporal dynamics, allowing the system to capture how the spectral properties change over time, which is particularly important for handling variations in speaking rates and articulation patterns. This finding aligns with previous findings (Snyder et al., 2018; Choi et al., 2015), where incorporating temporal features significantly improved speaker recognition performance under varied speaking conditions. In this study, the inclusion of MFCC-delta-delta improved accuracy from 60% to 62% in non-augmented data, further demonstrating the importance of modeling temporal fluctuations.

Formant frequencies, which represent the resonant frequencies of the vocal tract, provided additional information that enhanced speaker identification when combined with MFCCs, even though they underperformed as a standalone feature (59% accuracy without augmentation). While formants capture important physiological information about a speaker's vocal tract, they tend to be less reliable on their own, particularly when speech is affected by external factors such as noise or varying speaking rates (Hansen and Hasan, 2015). However, when paired with MFCCs, formants contribute valuable vocal tract information, improving the overall robustness of speaker recognition systems. This echoes findings from Nath and Kalita (Nath and Kalita, 2015), who demonstrated that combining formants with other features like MFCCs significantly enhanced speaker recognition accuracy, with results nearing 100% in some tasks. Similarly, Messaoud and Hamida (Messaoud and Hamida, 2011) found that integrating formant frequencies with MFCCs in a phone recognition system reduced the phone error rate by 3%, further validating the complementary nature of these features. These studies highlight how formants and MFCCs work together by covering different aspects of the speech signal, making the combined approach highly effective for speaker recognition.

Data augmentation played a pivotal role in enhancing system performance across all models. The most significant improvements were observed in the combined model, where accuracy increased from 70% to 75%, with similar gains in precision, recall, and F1-score. This finding is consistent with prior research, which demonstrated that augmentation methods such as time-stretching, pitch-shifting, and noise addition increase the diversity of the training data, enabling models to generalize better to unseen conditions (Nugroho et al., 2021; Ko et al., 2015). In our case, augmentation helped the model better handle variations in speaking rate and background noise, which are common in real-world applications. By exposing the model to these variations during training,

we effectively reduced overfitting, thus improving its ability to generalize to new data.

While MFCC-delta-delta and formant features saw improvements with augmentation, MFCCs alone also benefited, achieving a 65% accuracy compared to 60% without augmentation. This confirms that data augmentation is essential even when using robust features such as MFCCs, as it simulates real-world variability, making the model more resilient to changes in speech conditions (Koo et al., 2020). The increased accuracy of formant frequencies (68% with augmentation) suggests that, although formants alone may struggle with speaker discrimination in clean conditions, they become more useful when augmented, as they help capture subtle articulatory variations that may emerge under different speaking environments (Trottier et al., 2015).

5.1 Limitations and Future Work

A primary limitation of this study is the restricted dataset size and its language specificity. The dataset, consisting of only 18 male Persian speakers, limits the generalizability of the findings. Speaker verification systems often perform differently across languages due to phonetic and prosodic variations, raising uncertainty about the generalizability of these results to other linguistic contexts. Additionally, the small dataset may have constrained the model's ability to capture broader speaker variability. A more extensive and diverse dataset would be necessary to assess the system's robustness on a larger scale.

This study underscores the value of a hybrid acoustic feature approach for speaker identification, particularly when combined with data augmentation. The integration of MFCCs, delta features, and formant frequencies provided a multi-dimensional representation of vocal traits, enhancing performance under varied conditions. Future research could investigate additional feature combinations, such as prosodic features (e.g., intonation, rhythm) and voice quality measures (e.g., jitter, shimmer), to improve robustness, particularly for emotional speech or non-standard speaking styles.

While this study focused on traditional feature extraction methods, the use of deep learning-based embeddings—such as x-vectors(Snyder et al., 2018) or Transformer-based models(Vaswani, 2017)—holds significant potential. These models can learn speaker-specific characteristics directly from raw audio, reducing the reliance on manual feature engineering. Combining such approaches with advanced augmentation techniques could further enhance performance, enabling speaker identification systems to handle

challenging real-world conditions, such as noisy environments, emotional variability, and diverse speaking styles.

6 CONCLUSION

In this study, we investigated the resilience of various acoustic features in speaker identification across different speaking rates. The findings reveal a hierarchy of effectiveness among the examined parameters. The results indicate that vowel formant frequencies demonstrate a degree of resilience against changes in speaking rate, achieving an accuracy of 80 in speaker identification tasks. This suggests that while formant frequencies capture relevant speaker-specific information, their ability to distinguish between speakers may be somewhat compromised when faced with variations in speaking rate. Given the strong correlation between formant frequencies and the invariant physiological dimensions of the vocal tract, it is unsurprising that these frequencies remain relatively stable across varying speech rates. Despite the relatively stable vocal tract, dynamic articulatory adjustments required for different speaking rates could potentially introduce variability into formant measurements.

REFERENCES

Agwuele, A., Sussman, H. M., and Lindblom, B. (2009). The effect of speaking rate onconsonant vowel coarticulation. *Phonetica*, 65(4):194–209.

Alex, A., Wang, L., Gastaldo, P., and Cavallaro, A. (2023). Data augmentation for speech separation. *Speech Commun.*, 152:102949.

Anupama, V., Amrutha, C., Varshini, G. A., Nandan, G. S. G., and Vivek, G. S. S. (2022). A mfcc-cnn based voice authentication security. *Int. J. Eng. Technol. Manag. Sci.*, 4(6).

Asadi, H., Nourbakhsh, M., Sasani, F., and Dellwo, V. (2018). Examining long-term formant frequency as a forensic cue for speaker identification: An experiment on persian. In *Proc. 1st Int. Conf. Lab. Phon. Phonol.*, pages 21–28.

Bahari, M. H. and Van Hamme, H. (2011). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In *BIOMS*, pages 1–6. IEEE.

Boersma, P. and Weenink, D. (2021). Praat: Doing phonetics by computer [computer program](version 6.2. 22). *Retrieved fromwww. praat. org.*

Chakravarty, N. and Dua, M. (2023). Data augmentation and hybrid feature amalgamation to detect audio deep fake attacks. *Physica Scripta*, 98.

Choi, Y. H., Ban, S. M., Kim, K.-W., and Kim, H. S. (2015). Evaluation of frequency warping based fea-

tures and spectro-temporal features for speaker recognition. *Phonetics and Speech Sciences*, 7(1):3–10.

Corretge, R. (2022). Praat vocal toolkit. Available at: http://www.praatvocaltoolkit.com.

Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.*, 28(4):357–366.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.*, 19(4):788–798.

Gay, T., Ushijima, T., Hiroset, H., and Cooper, F. S. (1974). Effect of speaking rate on labial consonant-vowel articulation. *Journal of Phonetics*, 2(1):47–63.

Gold, E., French, P., and Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. In *Proc. Meet. Acoust.*, volume 19. AIP Publishing.

Hansen, J. H. and Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Process. Mag.*, 32(6):74–99.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97.

Imaizumi, S. and Kiritani, S. (1989). Effect of speaking rate on formant trajectories and inter-speaker variations. *Ann. Bull. RILP*, 23:27–37.

Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M. Z., and Ali, I. (2020). Text-independent speaker identification through feature fusion and deep neural network. *IEEE Access*, 8:32187–32202.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Interspeech*, volume 2015, page 3586.

Koo, H., Jeong, S., Yoon, S., and Kim, W. (2020). Development of speech emotion recognition algorithm using mfcc and prosody. *ICEIC*, pages 1–4.

Liu, X., Sahidullah, M., and Kinnunen, T. (2021). Learnable mfccs for speaker verification. In *ISCAS*, pages 1–5. IEEE.

Lounnas, K., Lichouri, M., and Abbas, M. (2022). Analysis of the effect of audio data augmentation techniques on phone digit recognition for algerian arabic dialect. *ICAASE*, pages 1–5.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies. *Int. J. Speech Lang. Law*, 13(1):89–126.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *SciPy*, pages 18–24.

Mefferd, A. S. and Green, J. R. (2010). Articulatory-to-acoustic relations in response to speaking rate and loudness manipulations. *J. Speech Lang. Hear. Res.*, 53:1206–1219.

Messaoud, Z. B. and Hamida, A. (2011). Combining formant frequency based on variable order lpc coding with acoustic features for timit phone recognition. *Int. J. Speech Technol.*, 14:393.

Nath, D. and Kalita, S. (2015). Composite feature selection method based on spoken word and speaker recognition. *Int. J. Comput. Appl.*, 121:18–23.

Nolan, F. (1987). The phonetic bases of speaker recognition: Cambridge studies in speech science and communication, cambridge university press, cambridge, 1983, 221 pp. isbn 0-521-24486-2.

Nugroho, K., Noersasongko, E., Purwanto, Muljono, and Setiadi, D. (2021). Enhanced indonesian ethnic speaker recognition using data augmentation deep neural network. *J. King Saud Univ. Comput. Inf. Sci.*, 34:4375–4384.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41.

Rose, P. (2002). *Forensic Speaker Identification*. International Forensic Science and Investigation. Taylor & Francis.

Shahrebabaki, A. S., Imran, A. S., Olfati, N., and Svendsen, T. (2018). Acoustic feature comparison for different speaking rates. In *Proc. Human-Computer Interaction (HCI)*, pages 176–189. Springer.

Shaiman, S., Adams, S. G., and Kimelman, M. D. (1997). Velocity profiles of lip protrusion across changes in speaking rate. *J. Speech Lang. Hear. Res.*, 40(1):144–158.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. *ICASSP 2018*, pages 5329–5333.

Tilsen, S. (2014). Selection and coordination of articulatory gestures in temporally constrained production. *Journal of Phonetics*, 44:26–46.

Trottier, L., Chaib-draa, B., and Giguere, P. (2015). Temporal feature selection for noisy speech recognition. In *Proc. Can. Conf. Artif. Intell.*, pages 155–166.

Tuller, B., Harris, K. S., and Kelso, J. S. (1982). Stress and rate: Differential transformations of articulation. *J. Acoust. Soc. Am.*, 71(6):1534–1543.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Weismer, G. and Berry, J. (2003). Effects of speaking rate on second formant trajectories of selected vocalic nuclei. *J. Acoust. Soc. Am.*, 113(6):3362–3378.

Xie, W., Nagrani, A., Chung, J. S., and Zisserman, A. (2019). Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019*, pages 5791–5795. IEEE.

Zeng, X., Yin, S., and Wang, D. (2015). Learning speech rate in speech recognition. *arXiv preprint arXiv:1506.00799*.

Zhou, Y., Xiong, C., and Socher, R. (2017). Improved regularization techniques for end-to-end speech recognition. *ArXiv*, abs/1712.07108.