https://jrl.ui.ac.ir/?lang=en

Journal of Researches in Linguistics
E-ISSN: 2322-3413
17(2), 29-38
Received: 03.02.2025 Accepted: 06.03.2025

Research Paper

Feature Selection for Kurdish-Persian Bilingual Speech: A Comparative Study of Formant Frequencies and MFCCs

Maral Asiaee
Experimental phonology and phonetics laboratory, Faculty of English, Adam Mickiewicz University, Poznań, Poland
marasi@amu.edu.pl

Homa Asadi*
Department of Linguistics University of Isfahan, Isfahan, Iran
h.asadi@fgn.ui.ac.ir

Abstract
This study examines the effectiveness of long-term formant frequency distributions (LTFDs), Mel-frequency cepstral coefficients (MFCCs), and their combined application in distinguishing Kurdish-Persian bilingual speakers. Speech samples were collected from 20 early male bilingual speakers who read the fable '*The North Wind and the Sun*' in Kurdish (Sorani dialect) and Persian. The Random Forest algorithm was employed to analyze the data. Feature importance for formant frequencies and MFCCs was evaluated using the mean decrease in accuracy metric. The results indicated that LTFD measures provided moderate accuracy in speaker differentiation, reflecting their capacity to capture vowel-related articulatory patterns. In contrast, MFCCs demonstrated superior performance, effectively encoding spectral and speaker-specific characteristics. When LTFDs and MFCCs were combined, system accuracy was slightly improved compared to using MFCCs alone. This marginal enhancement underscores the potential benefits of integrating LTFDs with MFCCs in forensic voice comparison, where even small gains can have significant practical implications. The findings contribute to a deeper understanding of bilingual speaker variability and provide insights for optimizing speaker identification systems in bilingual contexts.
Keywords: Bilingual speakers, Kurdish, Persian, long-term formant frequencies, Mel-frequency cepstral coefficients

1. Introduction

Forensic voice comparison (FVC) involves comparing an unknown criminal's sample to a known suspect(s)' sample(s). The main objective of FVC is to provide well-informed opinions and support law enforcement agencies and legal decision-makers, such as judges or juries, in determining whether the known and unknown voice samples originate from the same individual or different speakers (Chan & Wang, 2024; Gold, 2014). Expert forensic phoneticians achieve this goal by employing different methods, including acoustic analysis, auditory analysis, a combination of auditory and acoustic analysis, fully automatic speaker recognition (ASR), or human-assisted ASR. Results from several international surveys of FVC practices among law-enforcement agencies and forensic practitioners reveal that the most widely implemented approach involves integrating auditory and acoustic phonetic analysis (Gold & French, 2011, 2019; Morrison et al., 2016). However, ASR methods, whether fully automatic or human-assisted, were reported to be more prevalent in North America (Morrison et al., 2016).

In FVC, selecting robust acoustic parameters ensures reliable speaker identification. While implementing the auditory-acoustic method, expert forensic phoneticians must determine which acoustic parameters are more effective in distinguishing speakers. Consequently, over the past few decades, many research studies have focused on examining the role of various phonetic parameters, such as fundamental frequency, formant frequencies, and voice quality, in shaping a speaker's unique vocal signature. These parameters are believed to be reflective of articulatory settings, individual

*Corresponding author

https://doi.org/10.22108/jrl.2025.144241.1894

speech habits, and even socially acquired behaviors by speakers (Chan & Wang, 2024). Among these parameters, formant frequencies are reliable indicators of vowel quality and articulatory configurations. They are the acoustic correlates of vocal tract resonances and are primarily influenced by the shape and size of a speaker's supralaryngeal vocal tract, as they are inversely correlated with the speaker's vocal tract length.

Over time, researchers have employed various methods to analyze formant frequencies. The mid-point center frequency method, the most common approach, measures formants at vowel midpoints, representing their articulatory targets (Jessen, 2008; Rose, 2002). However, this approach overlooks dynamic format variations (Goldstein, 1976; McDougall & Nolan, 2007), leading to alternative techniques. Long-term spectra (LTS) analyze formant trajectories by averaging spectral slices but include voiceless portions and background noises, reducing the precision of the approach (Nolan & Grigoras, 2005). To address this, the long-term formant distribution (LTFD) method was introduced, focusing solely on vowels and voiced segments across entire speech samples rather than individual speech sounds, providing a more accurate representation of formant dynamics (Nolan, 1983; Rose, 2002; Gold, 2014).

Nolan and Grigoras (2005), who first proposed the LTFD method, claimed that LTFDs represent not only the physiological characteristics of an individual's vocal tract but also the speaker's overall articulatory habits. Empirical findings indicate that mean LTF1 values correlate with variations in laryngeal height, while elevated mean LTF2 values are associated with a more anterior positioning of the tongue body (Gold et al., 2013; Lo, 2021; Nolan & Grigoras, 2005).

While many studies have explored and reported the speaker-specificity of LTFDs in monolinguals (Asadi et al., 2018; Gold et al., 2013; McDougall, 2004; Moos, 2010; Skarnitzl et al., 2015), research on bilingual speakers has been limited and has yielded mixed results. Lo (2021) studied the language- and speaker-specificity of LTFDs in 60 male English-French bilinguals. He found systematic differences in LTFDs between the two languages, with French exhibiting higher LTF2-4 values than English, reflecting differences in vowel inventories and language-specific phonetic settings. However, a high degree of within-speaker consistency was also observed across the languages. He reported that while LTFDs were more effective for speaker discrimination in same-language comparisons, they still provided speaker-specific information across both languages. Asiaee et al. (2019) investigated the effectiveness of LTFDs in discriminating bilingual speakers, analyzing spontaneous speech samples from six Arabic-Persian bilinguals (three male, three female). They found LTF1 and LTF3 to be effective in discriminating bilinguals. In a study conducted by Cho and Munro (2017), Korean-English bilinguals retained the general shapes of their LTFDs across languages, yet they exhibited lower LTF2 peaks in Korean compared to English. Heeren et al. (2014) measured LTFs in Dutch-Turkish bilinguals and reported LTF2 and LTF3 to be comparable between languages when spoken by the same speaker, however, the shape of their LTF2 varied cross-linguistically, suggesting that while some aspects of LFTDs remain speaker-specific, others may be influenced by language. Indeed, Kinoshita (2001) noted that variations in phonological systems across languages can lead to differing results for the same acoustic parameters, emphasizing the language-specificity of some parameters.

Apart from the auditory-acoustic approach in FVC, ASR methods have also gained popularity among forensic practitioners, with Mel-frequency cepstral coefficients (MFCCs) being the most widely employed as input features (Hughes et al., 2023; Ashar et al., 2020; Nagaraja & Jayanna, 2014; Luengo et al., 2008; Nagaraja & Jayanna, 2014; Zhen et al., 2001). MFCCs are a set of features that represent the spectral characteristics of a sound signal. They are extracted by applying a linear cosine transform to the logarithm of the power spectrum, which is mapped onto a nonlinear mel scale (Davis & Mermelstein, 1980). Since MFCCs are derived from the mel-frequency cepstrum, where frequency bands are spaced according to the mel scale and not linearly-spaced frequency bands, they are more effective in modeling human auditory perception (Mistry & Kulkarni, 2013; Tirumala et al., 2017).

Contradictory results have been reported regarding improved speaker identification when acoustic parameters are integrated with MFCCs. While some studies indicated that incorporating acoustic features into MFCC-based systems does not always significantly enhance system performance, others suggested that acoustic features can indeed complement MFCC-based systems. In a study on 75 male Australian English speakers, Chan and Wang (2024) did not find any significant improvement in the system when integrating MFCCs with long-term acoustic features. Overall, in their study, MFCCs outperformed long-term acoustic features. Hughes et al. (2017) performed likelihood ratio-based testing using MFCCs and LTFDs. The fusion of MFCCs and LTFDs resulted in only marginal performance improvements over the baseline MFCC system, suggesting that these measures primarily encode similar speaker-specific information. Similarly, in a study conducted on the /iau/ tokens produced by 60 female speakers of Mandarin, no substantial improvement was yielded when integrating a formal-trajectory-based system with a baseline MFCC system (Zhang et al., 2013). However, Hughes et al. (2023) examined the hesitation marker *um* in Southern British English and found that integrating dynamic formant information significantly improved the MFCC-based system's performance.

Despite advances in forensic voice comparison (FVC) and speaker identification techniques, relatively little research has focused on bilinguals, particularly regarding the integration of MFCCs and LTFDs to assess system performance. Bilingualism adds complexity to such analyses, as speakers exhibit distinct acoustic profiles in their native (L1) and non-native (L2) languages, influenced by factors such as linguistic proficiency, phonological interference, and articulatory habits. Therefore, this study focuses of Kurdish-Persian bilinguals, providing a valuable case for examining acoustic variability across languages. The acoustic differences and distinct phonetic inventories of Kurdish and Persian create an optimal environment for investigating how acoustic features reflect speaker-specific characteristics. Two key

acoustic features are examined in this analysis: formant frequencies, which relate to articulatory mechanisms, and MFCCs, which capture broader spectral properties of speech. This study seeks to answer the following research questions:

1) To what extent can traditional formant frequency parameters effectively differentiate between Kurdish-Persian bilingual speakers?

2) To what extent can Mel-Frequency Cepstral Coefficients (MFCCs) effectively differentiate between Kurdish-Persian bilingual speakers?

3) To what extent can the combination of traditional formant frequency parameters and MFCCs effectively differentiate between Kurdish-Persian bilingual speakers?

These questions are investigated using an exploratory study conducted on a speech dataset of Kurdish-Persian bilingual speakers. Understanding the relative contributions of formant frequencies and MFCCs in bilingual contexts has significant implications for both theoretical linguistics and practical applications in speaker recognition systems. This knowledge can enhance the accuracy of speaker identification methodologies, particularly for diverse linguistic populations. Through the application of advanced machine learning models, we can better understand how individual acoustic profiles manifest in bilingual speech, ultimately advancing both forensic phonetics and our broader understanding of bilingual speech production.

2. Methodology

The subsequent sections detail participant information, audio data acquisition procedures, data preprocessing techniques, the methodology for extracting formant frequencies and MFFCs, and the selected statistical methods for data analysis.

2.1 Data collection

Speech samples were collected from 20 simultaneous male bilinguals of Kurdish-Persian. All participants spoke a Sorani dialect of Kurdish and had an age range of 25-39 years (Mean = 36, SD = ±4.76). Each participant read the fable "*The North Wind and the Sun*" once in Persian and once in Kurdish during separate recording sessions. Recordings were conducted using a ZOOM H5 hand-held recorder set at 44.1 kHz sampling rate and 16-bit resolution. The recorder was positioned 20 cm from the speaker's mouth at a 45-degree angle. All recordings were performed in a quiet room to minimize background noise. Participants were instructed to read at their natural speaking rate, pitch, and loudness.

2.2 Pre-processing the speech samples

Prior to acoustic analysis, all vowel segments within the speech signals were extracted using the "Extract Vowels" command within the Praat Vocal Toolkit (Corretge, 2022). This freely accessible plugin, integrated into the Praat software environment (Boersma & Weenink, 2022, version 6.2.09), offers a collection of automated scripts for diverse voice processing operations. After extraction, all isolated vowel segments were concatenated into a single continuous sequence for subsequent analysis.

2.3 Formant frequencies extraction

Following the preprocessing phase, formant frequencies were extracted, focusing on the first four formants (F1, F2, F3, and F4) of the concatenated vowel sequences. To capture the inherent temporal dynamics of formant frequency variations over time, a long-term analysis approach was adopted. This method has been demonstrated in previous research (Nolan, 1983; Rose, 2002; Nolan & Grigoras, 2005; Moos, 2010; Gold et al., 2013; Gold, 2014) to yield more robust and accurate representations of speaker-specific characteristics when compared to short-term analysis techniques. The extraction of formant frequencies was conducted using the LPC-Burg algorithm within the Praat software, with an LPC order of 12 and a ceiling frequency of 5000 Hz, ensuring accurate tracking of formant trajectories. The analysis was carried out with a frame size of 5 milliseconds, ensuring high temporal resolution and precision in the measurement of formant trajectories. To streamline this process, a pre-written script tailored explicitly for the Praat environment was utilized. This script automated the extraction process, reducing the potential for human error and ensuring consistency across all speech samples.

2.4 MFCC features extraction

The MFCC extraction process involved multiple stages. Initially, the speech signal was segmented into 15 ms frames with a 5 ms overlap. Each frame was subjected to a windowing function to minimize spectral leakage. Pre-emphasis filtering with a coefficient of 0.97 was applied to amplify high-frequency components, compensating for the natural attenuation of these frequencies in the human voice. The Fast Fourier Transform (FFT) was then applied to each frame to convert the time-domain signal into its frequency-domain representation. Subsequently, a Mel filter bank was employed to model the nonlinear frequency sensitivity of the human auditory system. The output of the Mel filters was logarithmically transformed to compress the dynamic range, enhancing the robustness of the feature set to variations in speech intensity. Finally, the Discrete Cosine Transform (DCT) was applied to the log-Mel energies to derive the MFCCs, which encapsulate the spectral characteristics of the speech signal in a compact and efficient representation. All processing steps were automated using a custom script within the Praat environment.

2.5 Statistical analysis

The Random Forest algorithm was employed for data analysis, given its effectiveness in handling high-dimensional and mixed-type data. The classification process was conducted independently for each feature set (long-term formant frequencies, MFCCs, and their combined set) for both Kurdish and Persian speech. To ensure a balanced representation of speakers and languages within the dataset, stratified sampling was employed to divide the data into training (70%) and testing (30%) subsets. A grid search approach, combined with 5-fold cross-validation, was utilized to optimize key hyperparameters of the Random Forest model, including the number of trees, maximum depth, and minimum samples per leaf. Model performance was assessed using classification accuracy, precision, recall, and F1-score to analyze classification errors. The feature importance scores computed by the Random Forest algorithm were analyzed to determine the contribution of individual features to speaker differentiation. Feature importance was calculated based on the mean decrease in accuracy, which measures the decrease in model performance when a specific feature is excluded or permuted. The feature with the largest mean decrease in accuracy is considered the most important, as removing or altering this feature leads to the largest drop in model performance. In contrast, features that cause little to no change in accuracy are deemed less important. For formant frequencies, the relative importance of F1, F2, F3, and F4 was evaluated to understand their role in encoding speaker-specific characteristics. For MFCCs, the contributions of each coefficient were analyzed to identify the most discriminative spectral features. For the combined feature set, the importance of formant frequencies and MFCC-derived features was compared to examine the interplay between these two feature types. The effectiveness of the three feature sets (Long-term formant frequencies, MFCCs, and combined features) was compared for Kurdish and Persian speech.

3. Results

3.1 Speaker classification using LTFD measures across Kurdish, Persian, and combined dataset

Table 1 presents the model performance metrics—accuracy, precision, recall, and F1-score—when LTF measures were used to differentiate bilingual speakers.

Table 1- Model performance using LTFD features in Persian, Kurdish, and combined data

Performance metrics	Kurdish	Persian	Combined
Accuracy	67.3%	60%	59.2%
Precision	66.7%	59.6%	58.9%
Recall	67.3%	60%	59.2%
F1-score	66.%	59.6%	58.8%

Results show that the model achieves 67.3% accuracy for Kurdish, indicating that 67.3% of predictions match the ground truth. Precision and recall are also high, resulting in an F1-score of 66.0%. These values suggest strong performance with balanced precision and recall. For the Persian dataset, performance is slightly lower, with an accuracy of 60% and a corresponding F1-score of 59.6%. These values indicate moderate performance with balanced precision and recall. The combined dataset shows the lowest performance, with an accuracy of 59.2% and an F1-score of 58.8%. The decrease may be attributed to increased variability in speaker characteristics across the two languages.

3.2 Speaker classification using MFCC measures across Kurdish, Persian, and combined dataset

Table 2 presents the performance of the model when MFCC features are used as features for speaker differentiation in Kurdish, Persian, and combined Kurdish-Persian dataset. The metrics reported include accuracy, precision, recall, and F1-score, which are closely aligned across all datasets.

Table 2- Model performance using MFCC features in Kurdish, Persian, and combined data

Performance metrics	Kurdish	Persian	Combined
Accuracy	91.8%	91.9%	90.9%
Precision	91.8%	91.8%	90.9%
Recall	91.8%	91.9%	90.9%
F1-score	91.8%	91.8%	90.9%

Based on the results, the model achieves an accuracy of 91.8% for the Kurdish dataset, indicating that nearly 92% of predictions correctly match the true speaker labels. Precision and recall are also 91.8%, respectively, reflecting the model's reliability and ability to retrieve speaker-specific instances effectively. The F1-score, which balances precision and recall, is similarly high at 91.8%, demonstrating overall robust performance. For the Persian dataset, the metrics are consistent with those of the Kurdish dataset, with accuracy, precision, recall, and F1-score all at 91.9%. This indicates that the model performs equally well in differentiating speakers within the Persian language. In the combined dataset, the model's accuracy is slightly lower at 90.9%. Precision, recall, and F1-score are also slightly reduced, each at 90.9%. These small decreases suggest that including both Persian and Kurdish speech introduces additional variability, making the task of speaker differentiation slightly more challenging. However, the results still indicate excellent performance and demonstrate the model's ability to generalize effectively across both languages.

https://doi.org/10.22108/jrl.2025.144241.1894

3.3 Speaker classification using a combination of LTFD and MFCC measures across Kurdish, Persian, and combined data

Table 3 shows the feature importance of the combination of LTFD measures and MFCCs across Kurdish and Persian, as well as the whole dataset for distinguishing speakers in Persian and Kurdish and their combined dataset.

Table 3- Model performance using LTF + MFCC features in Kurdish, Persian, and combined data

Performance metrics	Kurdish	Persian	Combined
Accuracy	92.5%	93.9%	92.3%
Precision	92.5%	93.9%	92.3%
Recall	92.5%	93.9%	92.3%
F1-score	92.4%	93.9%	92.3%

Results show that the highest classification performance was achieved for Persian speech, with an accuracy of 93.9%. For Kurdish speech, the metrics were slightly lower, with an accuracy of 92.5%. When the combined dataset was analyzed, the overall performance declined slightly, yielding an accuracy, precision, recall, and F1-score of 92.3% across all metrics. These findings suggest that speaker-specific characteristics may be more distinct within single-language datasets compared to the pooled dataset, where variability between the two languages may introduce additional complexity.

3.4 Feature importance based on LTFD measures

Table 4 illustrates the importance of four acoustic parameters (LTF1, LTF2, LTF3, and LTF4) for distinguishing speakers in Kurdish and Persian and their combined dataset. Each value represents the relative contribution of a feature to the differentiation task, normalized as percentages.

Table 4- Feature importance of LTFD parameters (LTF1–LTF4) across Kurdish, Persian, and combined data, represented as percentages

Feature	Kurdish (%)	Persian (%)	Combined (%)
LTF1	42.6	24.8	33.9
LTF2	38.7	22.0	30.2
LTF3	44.7	23.6	32.9
LTF4	42.5	21.5	30.9

Table 4 shows the relative importance of LTF measures for distinguishing speakers in Persian, Kurdish, and their combined dataset. The percentages indicate how much each feature contributes to speaker differentiation. In Kurdish, LTF3 and LTF1 emerge as the most significant, contributing 44.7`% and 44.7%. LTF4 and LTF2 also show substantial importance, contributing 42.5% and 38.7%, respectively. In Persian, the importance of the four LTF parameters is relatively balanced. LTF1 contributes the most with 24.8%, closely followed by LTF3 at 23.6% and LTF2 at 22.0%. LTF4 has the most minor contribution, accounting for 21.5%. This even distribution suggests that all four parameters play a similar role in speaker differentiation for Persian speakers. The combined dataset reflects an averaging of their contributions in Persian and Kurdish. LTF1 remains the most important parameter, contributing 33.9%, followed by LTF3 at 32.9%. LTF4 and LTF2 contribute 30.9% and 30.2%, respectively. The combined dataset maintains the prominence of LTF1 and LTF3, albeit with less variation compared to the individual datasets.

3.5 Feature importance based on MFCC measures

Table 5 displays the feature importance of MFCC parameters for speaker differentiation in Persian, Kurdish, and their combined dataset. The percentages represent each feature's relative contribution to speaker distinction.

Table 5- Feature importance of MFCC parameters across Kurdish, Persian, and combined data, represented as percentages

Feature	Kurdish (%)	Persian (%)	Combined (%)
C0	24.4	15.1	17.2
C1	11.3	11.1	9.0
C2	14.2	7.0	12.9
C3	16.1	7.7	11.5
C4	43.6	29.6	39.1
C5	20.9	17.3	16.4
C6	9.2	2.9	5.9
C7	26.0	24.1	29.6
C8	18.2	13.9	16.4
C9	6.2	3.0	4.8
C10	9.2	6.0	8.5
C11	5.3	1.7	3.3

Based on the results, C4 demonstrates the highest importance across all datasets, especially in Kurdish (43.6%) and the combined dataset (39.1%). In Persian, C4 also plays a significant role, contributing 29.6%. Other notable features include C7, which contributes 26.0% in Kurdish and 24.1% in Persian, and C0, which is influential in both Persian (15.1%) and Kurdish (24.4%). Lower contributions are observed for C11, C6, and C9, which collectively contribute less than 10% in all datasets. This indicates that these features are less relevant for distinguishing speakers. Overall, the results highlight the dominance of C4 and C7 in capturing speaker-specific characteristics, particularly in Kurdish speech. The combined dataset emphasizes C4 and C7 as the most important parameters for speaker differentiation.

3.6 Feature importance based on a combination of LTFD with MFCC measures across Kurdish, Persian, and combined data, represented as percentages

For the last analysis, we combined measures of LTFD with MFCCs to see how well the combination of these parameters works out in bilingual speaker identification. Figure 1 displays the bar chart showing the strengths of the selected parameters in showing between-speaker variability in bilingual speakers. As is evident in the chart, C4 and C7 had the best performance in Kurdish, Persian, and combined data. LTFD measures showed a moderate performance and F3 was the best parameter among them across Kurdish, Persian, and combined data. Moreover, LTF3 showed a better performance in Kurdish compared to the other two conditions.

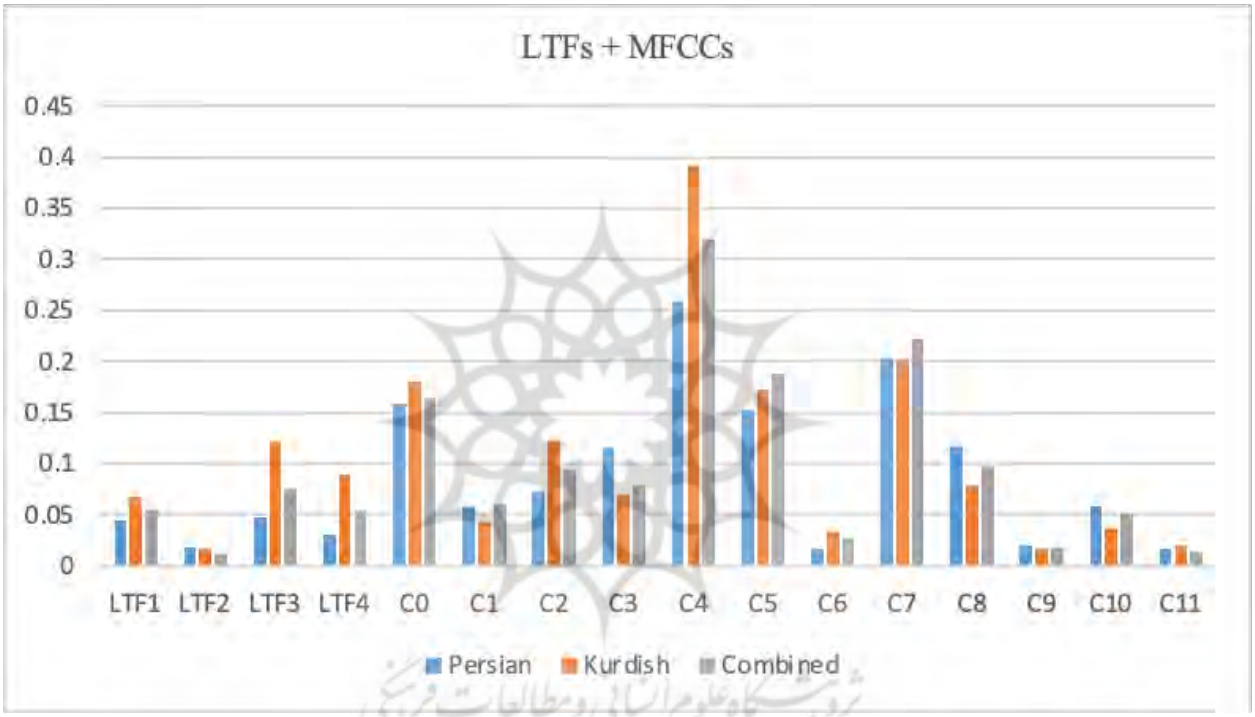

Figure 1- Feature importance of LTFD parameters (LTF1–LTF4) and MFCC features across Kurdish, Persian, and combined datasets, represented as percentages

4. Discussion

This study investigated the effectiveness of traditional formant frequencies and MFCCs and their combination in differentiating Kurdish-Persian bilingual speakers. Our primary objective was to determine which feature set exhibits the highest performance in capturing between-speaker variability within Kurdish-Persian bilingual speakers. We also examined model performance for each language individually and collectively to assess model accuracy and its variability across these conditions.

The results show that LTFD parameters capture speaker-specific information with varying degrees of success. For Persian, the model achieves a moderate accuracy of 60%. In contrast, the Kurdish dataset demonstrates significantly better performance, reaching an accuracy of 67.3%. This indicates that LTFDs are more effective in capturing speaker-specific features in a speaker's first language (Kurdish) compared to their second language (Persian). When speakers use their native language, they typically exhibit greater articulatory freedom, leading to increased variability in acoustic features such as formant frequencies. This enhanced flexibility may stem from their in-depth understanding of the native language's phonetic and phonological rules, allowing them to explore a wider range of articulatory gestures. This dynamic control subsequently contributes to greater individual variation, particularly in features like long-term formant frequencies, which capture unique speaker-specific characteristics. On the other hand, even early bilinguals may face constraints in their articulatory movements when speaking in a second language. These limitations can arise from incomplete mastery of the second language's phonetic system or a tendency toward first-language phonological interference. These factors can diminish the variability in formant frequencies observed in second-language speech,

https://doi.org/10.22108/jrl.2025.144241.1894

potentially obscuring speaker-specific traits that are more evident in the first language. Furthermore, Kurdish and Persian have distinct phonetic characteristics, and early bilinguals may exhibit more natural and individualized articulation patterns in their native language. This could enhance the variability captured by LTFDs, making them more effective for speaker identification in the first language. Moreover, the results for the combined dataset yielded the lowest performance, with an accuracy of 59.2%. The reduced performance likely results from increased variability in speaker characteristics across the two languages, which may introduce additional complexities when modeling speaker-specific information.

The analysis of feature importance in the Kurdish dataset showed that among LTFDs, LTF3 (44.7%) and LTF1 (42.6%) emerged as the most important features, highlighting their important role in capturing speaker-specific information. The Persian dataset shows balanced contributions from all four parameters, with LTF1 (24.8%) and LTF3 (23.6%) being slightly more prominent. The combined dataset reflects an averaging effect, with LTF1 (33.9%) and LTF3 (32.9%) remaining the most significant. The sensitivity of these formants to subtle variations in vocal tract shape likely contributes significantly to their effectiveness in distinguishing between individual speakers. For instance, F1 is closely associated with the degree of oral cavity aperture during vowel production while F3 relates to lip rounding and tongue positioning. The dominance of LTF1 and LTF3 in the analysis reinforces their utility in speaker identification tasks, particularly in bilingual and multilingual contexts. The findings support previous research (Asadi et al., 2023; Asadi & Alinezhad, 2020; Asiaee et al., 2019; Becker et al., 2008; Gold et al., 2013; He et al., 2019; Lo, 2021; Vaňková & Skarnitzl, 2014) and contribute to the growing evidence that these formants effectively capture speaker-specific variations. Furthermore, the results underscore the importance of considering language-specific characteristics when designing speaker recognition systems, as the prominence of certain features can vary based on the linguistic and acoustic properties of the speech data.

Regarding MFCCs, the model demonstrates a better performance in distinguishing speakers. The model achieves high accuracy for both Persian and Kurdish datasets (91.9% and 91.8%, respectively). The effectiveness of MFCCs in retrieving speaker-specific information further supports the established reliability of MFCCs. Previous research has also consistently demonstrated that MFCCs effectively capture between-speaker variability (Ashar et al., 2020; Leu & Lin, 2017; Liu et al., 2018; Luengo et al., 2008; Nagaraja & Jayanna, 2014; Zhen et al., 2001). Furthermore, the consistent performance across the two languages suggests that MFCCs effectively capture speaker variability, regardless of language-specific features. For the combined dataset, the model's accuracy slightly decreases to 90.9%, with corresponding reductions in precision, recall, and F1-score. This small decline reflects the additional variability introduced by combining Persian and Kurdish speech. Nonetheless, the overall performance remains excellent, demonstrating the robustness of MFCCs in distinguishing speakers across languages. These results underscore the capability of MFCCs to generalize effectively in bilingual contexts. As for the feature importance of MFCCs, C4 consistently demonstrates the highest importance across all datasets, especially in Kurdish (43.6%) and the combined dataset (39.1%). C7 and C0 also show notable contributions, particularly in Kurdish speech. The prominent role of C4 and C7 in capturing speaker-specific characteristics suggests that these cepstral coefficients are robust against within-speaker variability, potentially including variations arising from linguistic differences. C4, the fourth MFCC coefficient, reflects mid-range spectral features, representing the shape of the spectral envelope in the middle-frequency range. This coefficient is especially sensitive to vowel sounds, as vowels are characterized by distinct formant structures and vocal tract resonances, which vary uniquely across speakers. C7, the seventh MFCC coefficient, captures higher frequency details of the spectral envelope and highlights finer spectral variations. These variations can reveal subtle articulatory differences during vowel production, making C7 essential for identifying speaker-specific information embedded in speech signals.

In addressing how LTFDs and MFCCs perform together for bilingual speaker identification, results showed that the combined system achieved high performance across all datasets, with a slight improvement over MFCCs used independently. The model maintains an accuracy of around 92.5% and 93.9% for Kurdish and Persian data. The combined dataset shows a marginal decrease in performance, with accuracy at 92.3%. These results indicate that while the combination of LTFDs and MFCCs enhances the richness of feature representation, it does not significantly outperform MFCCs alone. Our findings diverge from those of Chan and Wang (2024), who reported no significant performance improvement and even a slight degradation when incorporating long-term phonetic features into an MFCC-based speaker identification system. In contrast, our study observed a very slight performance enhancement with the addition of LTFDs. While this improvement may seem marginal, it still holds potential value in forensic voice comparison tasks where even minor gains in accuracy can have significant implications.

5. Conclusion

This study investigated the effectiveness of LTFDs, MFCCs, and their combination for speaker differentiation in Kurdish-Persian bilinguals. The primary objective was identifying the most effective acoustic features for capturing between-speaker variability. We also evaluated model performance for each language individually and collectively. Results demonstrated that while LTFDs exhibit moderate discriminative power, their performance varies significantly across languages. In contrast, MFCCs consistently demonstrated robust speaker differentiation across both languages, proving effective in this bilingual context. Although combining LTFDs and MFCCs resulted in a slight performance enhancement over MFCCs alone, the improvement was marginal. While our findings demonstrate promising results, certain limitations of this study should be acknowledged. Firstly, the dataset exclusively comprises male speakers and is

relatively small, limiting the findings' generalizability. Future research should focus on expanding the dataset to include female voices and incorporate spontaneous speech data to better reflect natural speaking styles. Furthermore, future studies could explore more sophisticated feature fusion techniques, such as deep learning-based approaches. Additionally, incorporating other acoustic features, including prosodic features and voice quality measures, could potentially enhance speaker differentiation accuracy in bilingual settings.

References

Asadi, H., & Alinezhad, B. (2020). Speaker-specific features of simple vowels in Persian based on the source-filter theory. *Journal of Researches in Linguistics*, *12*(2), 241-262. [In Persian]. https://doi.org/10.22108/jrl.2021.128697.1577

Asadi, H., Asiaee, M., & Alinezhad, B. (2023). Acoustic analysis of parameters affecting the between-speaker variability in Persian-English bilinguals. *ZABANPAZHUHI (Journal of Language Research) 15*(47), 131-155. [In Persian]. https://doi.org/10.22051/jlr.2022.40224.2174

Asadi, H., Nourbakhsh, M., Sasani, F., & Dellwo, V. (2018). Examining long-term formant frequency as a forensic cue for speaker identification : An experiment on Persian. In M. Nourbakhsh, H. Asadi, & M. Asiaee (Eds.), *Proceedings of the First International Conference on Laboratory Phonetics and Phonology* (pp. 21–28). Neveesh Parsi Publications.

Ashar, A., Bhatti, M. S., & Mushtaq, U. (2020). Speaker Identification Using a Hybrid CNN-MFCC Approach. *2020 International Conference on Emerging Trends in Smart Technologies, ICETST 2020*. https://doi.org/10.1109/ICETST49965.2020.9080730

Asiaee, M., Nourbakhsh, M., & Skarnitzl, R. (2019). Can LTF discriminate bilingual speakers? *Proceedings of the 28th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*, 41–42.

Becker, T., Jessen, M., & Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian mixture models. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, January*, 1505–1508.

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer* (6.2.09). https://www.fon.hum.uva.nl/praat/

Chan, R. K. W., & Wang, B. X. (2024). Do long-term acoustic-phonetic features and mel-frequency cepstral coefficients provide complementary speaker-specific information for forensic voice comparison? *Forensic Science International*, *363*(August), 112199. https://doi.org/10.1016/j.forsciint.2024.112199

Cho, S., & Munro, M. J. (2017). F0, long-term formants and LTAS in Korean-English Bilinguals. *The 31st General Meeting Phonetic Soc. Japan*, 188–193.

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

Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366.

Gold, E. (2014). *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. The University of York.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech Language and the Law*, *18*(2), 293–307. https://doi.org/10.1558/ijsll.v18i2.293

Gold, E., & French, P. (2019). International Practices in Forensic Speaker Comparisons: Second Survey. *International Journal of Speech Language and the Law*, *26*(1), 1–20. https://doi.org/10.1558/ijsll.38028

Gold, E., French, P., & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proceedings of Meetings on Acoustics*, *19*. https://doi.org/10.1121/1.4800285

Goldstein, U. G. (1976). Speaker- identifying features based on formant tracks. *The Journal of the Acoustical Society of America*, *59*(1), 176–182. https://doi.org/10.1121/1.380837

He, L., Zhang, Y., & Dellwo, V. (2019). Between-speaker variability and temporal organization of the first formant. *The Journal of the Acoustical Society of America*, *145*(3), EL209–EL214. https://doi.org/10.1121/1.5093450

Heeren, W., Vloed, D., & Vermeulen, J. (2014). Exploring long-term formants in bilingual speakers. *IAFPA's 2014 Annual Conference Book of Abstracts*, *2014*, *3*, 39.

Hughes, V., Cardoso, A., Foulkes, P., French, P., Gully, A., & Harrison, P. (2023). Speaker-specificity in speech production: The contribution of source and filter. *Journal of Phonetics*, *97*. https://doi.org/10.1016/j.wocn.2023.101224

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & Segundo, E. S. (2017). Mapping across feature spaces in forensic voice comparison: The contribution of auditory-based voice quality to (semi-)automatic system testing. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2017-August*, 3892–3896. https://doi.org/10.21437/Interspeech.2017-1508

Jessen, M. (2008). Forensic Phonetics. *Language and Linguistics Compass*, *2*(4), 671–711. https://doi.org/10.1111/j.1749-818X.2008.00066.x

Kinoshita, Y. (2001). *Testing Realistic Forensic Speaker A Likelihood Ratio Based Approach Using Formants* [The Australian National University]. https://doi.org/10.1558/ijsll.v9i1.133

Leu, F. Y., & Lin, G. L. (2017). An MFCC-based speaker identification system. *Proceedings - International*

Conference on Advanced Information Networking and Applications, AINA, 1055–1062. https://doi.org/10.1109/AINA.2017.130

Liu, J. C., Leu, F. Y., Lin, G. L., & Susanto, H. (2018). An MFCC-based text-independent speaker identification system for access control. *Concurrency and Computation: Practice and Experience*, *30*(2), 1–16. https://doi.org/10.1002/cpe.4255

Lo, J. J. H. (2021). *Cross-Linguistic Speaker Individuality of Long-Term Formant Distributions: Phonetic and Forensic Perspectives*. 416–420. https://doi.org/10.21437/interspeech.2021-1699

Luengo, I., Navas, E., Sainz, I., Saratxaga, I., Sanchez, J., Odriozola, I., & Hernaez, I. (2008). Text independent Speaker identification in multilingual environments. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, *2004*, 1814–1817.

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aI/. *International Journal of Speech, Language and the Law*, *11*(1), 103–130. https://doi.org/10.1558/sll.2004.11.1.103

McDougall, K., & Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS XVI)* (pp. 1825–1828). Universität des Saarlandes.

Mistry, D. S., & Kulkarni, P. A. V. (2013). Overview: Speech Recognition Technology, Mel- frequency Cepstral Coefficients (MFCC), Artificial Neural Network (ANN). *International Journal of Engineering Research & Technology*, *2*(10), 1994–2002.

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician*, *101/102*, 7–24.

Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Goemans Dorny, C. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International*, *263*, 92–100. https://doi.org/10.1016/j.forsciint.2016.03.044

Nagaraja, B. G., & Jayanna, H. S. (2014). Efficient window for monolingual and crosslingual speaker identification using MFCC. *ICACCS 2013 - Proceedings of the 2013 International Conference on Advanced Computing and Communication Systems: Bringing to the Table, Futuristic Technologies from Around the Globe*, 19–22. https://doi.org/10.1109/ICACCS.2013.6938702

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge University Press.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, *12*(2), 143–173. https://doi.org/10.1558/sll.2005.12.2.143

Rose, P. (2002). *Forensic Speaker Identification*. Taylor and Francis.

Skarnitzl, R., Vaňková, J., & Nechanský, T. (2015). Speaker discrimination using formant trajectories from casework recordings: Can LDA do it? *Proceedings of the 18th International Congress of Phonetic Sciences*, 1–5.

Tirumala, S., Shahamiri, S., Garhwal, A., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, *90*, 250–271. https://doi.org/10.1016/J.ESWA.2017.08.015

Vaňková, J., & Skarnitzl, R. (2014). Within- and between-speaker variability of parameters expressing short-term voice quality. *Proceedings of the International Conference on Speech Prosody*, *September*, 1081–1085. https://doi.org/10.13140/2.1.2377.1520

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison-Female voices. *Speech Communication*, *55*(6), 796–813. https://doi.org/10.1016/j.specom.2013.01.011

Zhen, B., Xihong, W., Zhimin, L., & Huisheng, C. (2001). On the Importance of Components of the MFCC in Speech and Speaker Recognition. *Acta Scientiarum Naturalium-Universitatis Pekinensis*, *37*(3), 371–378. https://doi.org/10.21437/ICSLP.2000-313

https://doi.org/10.22108/jrl.2025.144241.1894