# Formants and Prosodic Features' Effects on Arabic Speaker Identification Accuracy in Noisy Environments

**Khadidja Nesrine Boubakeur***
Laboratory of Spoken Communication and Signal Processing, USTHB, Algiers, Algeria
Scientific and Technical Research Center for the Development of Arabic Language CRSTDLA, Algiers, Algeria.
*Email: boubakeur.khadidja@gmail.com*

**Mohamed Debyeche**
Laboratory of Spoken Communication and Signal Processing, USTHB, Algiers, Algeria.
*Email: mdebyeche@gmail.com*

**Received:** 04/12/2024 **Accepted:** 19/12/2024 **Published:** 30/12/2024

**Abstract:**

This study investigates the use of formants and prosodic features, specifically pitch and intensity, for speaker identification in real conditions. To enhance the robustness of the acoustic models against speech signal variations in noisy environments, Mel-Frequency Cepstral Coefficient (MFCC) are added to these features. A Speaker Identification system based on Hidden Markov Models (HMM) is implemented in the independent text mode. The combination of formants and prosodic features with cepstral features improves the identification accuracy, particularly in high-noise environments, up to 10%, in comparison to an MFCC based system. The results show that the use of multivariate feature vectors significantly improves the performance of an identification system in the presence of noise compared to an MFCC-based system.

***Keywords***: ASI – MFCC – Prosodic features – Pitch – Energy – Formants – HMM - Noise.

* Corresponding author: Khadidja Nesrine Boubakeur

# تأثيرالبو اني الصوتية والخصائص النبرية على دقة التعرف على المتكلم العربي في بيئات صاخبة.

## الملخص:

تتناول هذه الدراسة استخدام البواني الصوتية والخصائص النبرية، تحديدًا النغمة والشدة، لتحديد هوية المتحدث في بيئات صاخبة. لتعزيز قوة النماذج الصوتية ضد التغيرات في إشارات الكلام في بيئات صاخبة، تمت إضافة معاملات (MFCC) إلى هذه الميزات. تم انشاء نظام تعرف آلي على المتكلم يعتمد على نماذج ماركوف المخفية(HMM) . أظهر الجمع بين البواني الصوتية والميزات النبرية مع معاملات (MFCC) تحسنًا في دقة التعرف على المتكلم، خاصةً في البيئات ذات الضوضاء العالية، بنسبة تصل إلى 10% مقارنة بالنظام القائم فقط علىMFCC . تُظهر النتائج أن استخدام معاملات متعددة يعزز بشكل كبير من أداء نظام التعرف الآلي على المتكلم في وجود الضوضاء، مقارنةً بالنظام القائم على MFCC وحده.

**كلمات مفتاحية:** البواني الصوتية- الخصائص النبرية- النغمة- الشدة- التعرف الآلي على المتكلم.

# Effets des formants et des paramètres prosodiques sur la précision de l'identification du locuteur arabe dans des environnements bruités.

## Résumé :

Ce travail porte sur l'utilisation des formants et des paramètres prosodiques, notamment la fréquence fondamentale (Pitch) et l'intensité, pour l'identification du locuteur dans un environnement bruité. Afin d’améliorer la robustesse des modèles acoustiques face aux variations du signal de parole dans des environnements bruités, les paramètres cepstraux de fréquence Mel (MFCC) sont ajoutés à ces caractéristiques. Un système d'identification automatique du locuteur basé sur des modèles de Markov cachés (HMM) est mis en œuvre. La combinaison des formants et des paramètres prosodiques avec les caractéristiques cepstrales permet d'améliorer la robustesse des systèmes d'identification, en particulier dans des environnements très bruyants, avec une amélioration de 10 % par rapport à un système basé sur les MFCC. Les résultats montrent que l'utilisation de vecteurs de caractéristiques multivariées permet l'amélioration des performances d'un système d'identification en présence de bruit par rapport à un système basé sur les MFCC.

***Mots clés* :** Formants – Paramètres prosodiques – Fréquence fondamentale - Intensité – Identification Automatique du Locuteur.

## INTRODUCTION

Speech is the basic way of communication for humans. In addition to transmitting the message, a speech signal also indicates the speaker's identity (Doddington, 1985). The process of identifying speakers involves distinguishing them according to their vocal variances, which include Physiological differences and variations in speaking (Doddington, 1985; Rabiner & Juang, 1993). Mel-Frequency Cepstral Coefficients (MFCCs) are commonly used in speaker identification systems to describe speech signal (Huang, Acero, Hon, & Reddy, 2001), as they capture specific spectral features that can be used to differentiate speakers.

Speaker identification in noisy environments is challenging because background noise can affect the speech signal, making it difficult to extract speaker features. According to the literature, machine learning-based systems using MFCC coefficients as parameterisation have not performed well under these conditions (Huang et al., 2012). To overcome these limitations, researchers often combine MFCC with other features (Arinaitwe et al., 2024) or apply advanced methods such as feature enhancement (Al-Karawi & Mohammed, 2021), noise reduction (Arinaitwe et al., 2024), or hybrid models (Arinaitwe et al., 2024) to improve performance in noisy environments.

Linguistically motivated features can be useful for robustness in speaker identification (Kreiman & Sidtis, 2011; McDougall, 2006; Reynolds & Rose, 1995), so a number of methods and approaches have been proposed (Kreiman & Sidtis, 2011; McDougall, 2006; Reynolds & Rose, 1995), such as prosodic features (Singh, Khan, & Shree, 2012; Mary & Yegnanarayana, 2006) and formants (Boubakeur et al., 2022; McDougall, 2006; Falek et al., 2011). The combination of MFCCs with formant and prosodic parameters permits us to take advantage of the power of both sorts of characteristics to increase the performance of the identification system in difficult environments (Droua-Hamdani, 2020; Amrous & Debyeche, 2012).

In a previous work (Boubakeur et al., 2022), we studied the impact of combining MFCC with pitch and energy on a speaker identification system in noisy environment. A gain of 9% was obtained for a highly noisy environment (0db). In this study, we investigate the contribution of formants and prosodic features to speaker identification in both clean and noisy speech signals. Pitch, Energy and the first four formants are combined with MFCC parameters and integrated in an HMM-based speaker identification system.

# 1. HMM BASED SPEAKER IDENTIFICATION SYSTEM

HMM is a commonly used statistical technique in speaker recognition (Amrous et al., 2011) due to its expressive power in process modelling (Ferrer et al., 2010; Amrouche et al., 2019). HMM-based speaker identification systems use the ability of HMMs to model temporal sequences and capture the dynamic nature of speech to accurately identify speakers based on their voices (Ji and al., 2015). As illustrated in Figure 1, an Hmm-based speaker identification system has three main components: feature extraction, model training and identification. A speech pre-processing step is required before feature extraction.

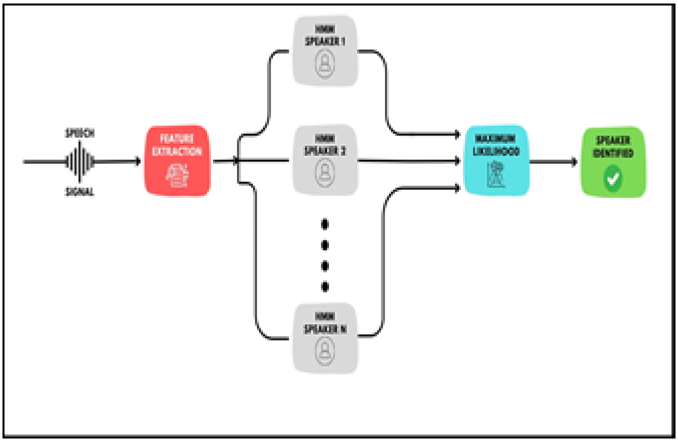

**Figure 1. HMM-based speaker identification system.**

## *A. Pre-processing*

Preprocessing prepares the speech data for efficient feature extraction and modelling, making it a critical step in the identification process. Common preprocessing steps are:

- Sampling: This involves converting the continuous audio signal into a discrete digital representation that can be analysed by the system.
- Pre-emphasis: This consists of applying a pre-emphasis filter to boost high frequency components in order to compensate for the effects of speech production that attenuate higher frequencies.”
- Segmentation: involves dividing the speech signal into smaller,

overlapping frames of 20 to 30 ms to capture short-term spectral properties.

- Windowing: consists of applying windowing functions such as Hamming to minimise spectral leakage in order to reduce edge effects when analysing speech segments.

### *B. Feature Extraction*

Feature extraction consists of transforming the speech signal into a set of features representing speaker's characteristics. These features are then used for training and identification.

### *C. Model training*

During the training phase, each speaker is assigned a Hidden Markov Model (HMM) (see Figure 1). Each HMM is specified by parameters such as transition probabilities (the likelihood of transitioning from one state to another) and emission probabilities (the likelihood of seeing a particular feature given a state). The Viterbi algorithm (Cui & Chen, 2010) is used to initialise the HMM models, followed by the Baum-Welch algorithm (Cui & Chen, 2010) to refine the training process.

### *D. Identification*

The identification procedure involves determining the likelihood of the observed sequences (the speaker to be identified) and the acoustic models. The speaker is identified by comparing the observed sequence with the acoustic model with the highest likelihood. This likelihood is computed using the Viterbi technique (Cui & Chen, 2010).

## 2. SYSTEM FEATURES

MFCC, formants, and prosodic features are the parameters used in this work. Some of their theoretical foundation is described in the following sections:

### *A. MFCC*

MFCC are one of the most popular features in speaker recognition systems (Tiwari, 2010). They capture a spoken signal's short-term power spectrum and are intended to mimic how the human ear hears sound (Leu & Lin, 2017). In clean surroundings, MFCC outperform other recognition algorithms (Arinaitwe et al., 2024). However, their performance suffers dramatically in noisy conditions. To solve this issue, the authors propose integrating MFCCs with formants and prosodic characteristics. The MFCC coefficients are calculated by taking the linear cosine transform of a logarithmic power spectrum on a non-linear Mel frequency scale. The MFCC coefficients for each analysis window are derived using equation (1), as follows :

$$MFCC = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi n\left(m+\frac{1}{2}\right)}{M}\right), \quad 0 \leq n \leq M \tag{1}$$

M is the number of channels in the filter bank, and E[m] is the energy of a given filter.

### *B. Formants*

Formants, or the resonant frequencies of the vocal tract, are critical to identifying speakers because they can reveal details about how the speaker's vocal tract works. Unlike MFCC, formants are less affected by noise, so adding formant frequencies can help distinguish speakers even in the presence of noise.

A formant in a speech wave is an acoustic energy concentration centered around a specific frequency. There are multiple formants, with varying frequencies. F1 and F2 both reflect the phonetic quality of a vowel. The higher formants, F3, F4, and F5, convey more speaker-specific traits (Fairclough et al., 2023). In this work, we have chosen to employ the first four formants' frequencies. They are calculated using the LPC spectrum model's maxima, which are the complex roots of the subsequent polynomial:

$$1 + \sum_{i=1}^{P} a_i z^i = 0 \tag{2}$$

The Lpc order is denoted by P.

Only the roots with positive imaginary parts are retained. The angles of these roots represent the formant frequencies, which are then transformed from radians per sample to Hertz. The formant bandwidths are determined from the distance of the roots from the unit circle.

### *C. Prosodic features*

Prosodic features, such as pitch, energy, duration, speech rate and intonation, represent learned speech habits and characteristics and provide complementary information that improves the accuracy and robustness of speaker identification systems. These features are generally more robust to background noise than MFCC. System performance can be improved by combining prosodic and spectral features, especially under difficult conditions. Our study focuses on pitch and energy.

*1) Pitch:* it is an important aspect of speech signals as it represents the fundamental frequency and captures the intonation of the speaker. It is defined as the frequency at which the vocal folds vibrate as air flows over the glottis. This feature is required for speech analysis and is estimated using the autocorrelation function. For a speech window {s(n), n = 0, 1, ... , Ns - 1} the autocorrelation function is defined as:

$$R(k) = \frac{1}{N}\sum_{n=0}^{N_s=1-k} s(n)s(n+k),\ k = 0, \dots., N_s - 1 \qquad (3)$$

Where Ns is the number of autocorrelation points to be computed

*2) Energy:* Reflects the loudness or intensity of the speech signal. It captures variations in speech amplitude, which can reveal details about the speaker's articulation and speech patterns (Singh & Khan, 2015). Short segments, or frames, of the speech signal are used to calculate energy. The short-term energy (St) t=1, T of a sampled signal over a window of length T is defined by the following equations:

$$E = \frac{1}{T}\sum_{t=1}^{T} s_t^2 \qquad (4)$$

## 3. EXPERIMENTS AND RESULTS

### *A. Speech Corpus*

The ARADIGITS database used in this work is a collection of spoken Arabic numerals (0-9) recorded by 60 Algerian speakers with different regional accents. The recordings, made in a clean environment (ambient noise levels below 35 decibels). The training set contains 1440 speech files from 60 individuals (30 male and 30 female), each of which repeats the digits 0-7 three times. The test database contains 360 speech files from the same 60 speakers, each repeating the numbers 8 and 9 three times.

### *B. Experiments Evaluation*

Our goal is to enhance speaker identification in noisy environment by integrating the MFCC with formants and prosodic features. For this purpose, A text-independent automatic speaker identification system based on HMM has been put into place. we first describe the baseline system and then show the outcomes of the fusion system.

*1) Baseline system:* is an HMM-based, text-independent speaker identification system built using HTK (Hidden Markov Toolkit) software. (Young, Odell, et al. 2002). The system is designed based on the following choices:

- *Signal Representation:* The acoustic signal is represented using 12 MFCCs, along with their first- and second-order differential coefficients, resulting in a total of 39 coefficients.

- *Speaker Modeling:* Each speaker is modeled using a Hidden Markov Model (HMM).

- *Acoustic Model Topology:* All HMMs adopt a Bakis topology with three emitting states.

*2) Emission Probability:* A linear combination of eight Gaussian distributions with diagonal covariance matrices is used to simulate the emission probability for each state.

*3) Proposed system:* Pitch (F0), energy (E) and the first four formants (F1, F2, F3, F4) are extracted using the Praat package (Boersma, 2006) based on methods described in section 3. To create a singular observation at the ASI system's input, these features are combined with their first (Δ) and second (ΔΔ) derivatives to create a unique vector.

*3) Experiments results:*

We investigated the robustness of an Automatic speaker identification system in a quiet and noisy environment. Babble speech and manufacturing noise are the two forms of noise that we have used to corrupt the database in order to create adverse situations. To obtain a signal-to-noise ratio (SNR) of 10 dB, 5 dB, and 0 dB, both noises were taken from the NOISEX92 database (Varga & Steeneken, 1993) and added to the speech signal. The acoustic models are estimated using a clean speech database, with noises introduced during the test phase. Table 1 shows the speaker identification rates obtained with both systems: the baseline system and the proposed system (which uses MFCC with formants and prosodic characteristics).

These rates are shown in both clean and noisy conditions. Additionally, a third system has been implemented, in which the acoustic vectors are based on the fusion of MFCC with prosodic features. The identification rate is expressed as a percentage and is defined as follows:

$$IR(\%) = \frac{Number\ of\ tests\ correctly\ identified}{Total\ number\ of\ tests} * 100 \tag{5}$$

TABLE 1.COMPARATIVE SPEAKER IDENTIFICATION RATES

| Environment | IR (%) | | | |
|---|---|---|---|---|
| | SNR (dB) | Baseline system | HMM (MFCC + Pros) | HMM (MFCC +Pros+ 4 Formants) |
| Clean | | 95.00 | 78.33 | 68.89 |
| Babble speech Noise | 10 | 57.78 | 52.5 | 48.06 |
| | 5 | 31.11 | 33.8 | 37.70 |
| | 0 | 14.72 | 17.7 | 18.61 |
| Factory Noise | 10 | 68.06 | 52.5 | 58.89 |
| | 5 | 41.39 | 40.83 | 45.56 |
| | 0 | 20.28 | 29.1 | 30.83 |

## 4. DISCUSSION

We note that the addition of noise to our system significantly degrades the recognition rate in a quiet environment (95% vs. 14.72%). The integration of prosodic parameters and formants does not bring any improvement in a quiet environment (95% vs. 68.89%) or in a low-noise environment (57.78% vs. 48.06%). This is due to the combination of different features in the same vector, which may not be ideal for clean environments, resulting in suboptimal performance. However, there was an improvement of 4% for babble speech noise at 0db (14.75% vs 18.61%) and a gain of 10% for factory noise at 0db (20.28% vs 30.83%).

## 5. CONCLUSION

In this study, we have developed an HMM-based using MFCC features as an acoustic model of the speaker's representative speech signal, the speaker identification system This system is considered to be the baseline system. Subsequently, a straightforward concatenation was employed to incorporate complementary parameters, namely the first four formants and prosodic features, specifically pitch and energy, in this instance. The reference system and the created systems' performances were assessed in identical circumstances. The results indicate that the use of formants and prosodic characteristics in the developed system significantly improves identification rates in real-world settings. In fact, in noisy conditions, the inclusion of these features leads to an approximately 10% improvement in accuracy compared to the baseline system under noisy conditions.

The information yielded by both formant and prosodic characteristics can be considered to be supplementary to that provided by cepstral features (MFCC). Formants are indispensable tools for acquiring knowledge about the physiological characteristics of a speaker's voice, as they capture the distinctive resonant frequencies of each individual's vocal tract. Pitch and energy are examples of prosodic qualities that offer a greater degree of information by expressing the speaker's speech patterns and intonation. By integrating these complementary feature sets, the speaker identification system can leverage the strengths of both spectral and temporal characteristics of speech. This multidimensional method allows for the representation of speaker-specific qualities with greater precision, thereby enhancing the ability to distinguish between speakers, particularly in challenging situations such as background noise and diverse speaking styles.

In conclusion, the combination of formants and prosodic characteristics offers a robust basis for developing more accurate and reliable speaker identification systems.

We suggest using articulatory features as complementary features and combining deep learning techniques with conventional feature extraction techniques as perspectives for this study.

## References

Al-Karawi, K. A., & Mohammed, D. Y. (2021). Improving short utterance speaker verification by combining MFCC and entropy in noisy conditions. *Multimedia Tools and Applications, 80*(14), 22231–22249.

Amrous, A. I., & Debyeche, M. (2012). Robust Arabic multi-stream speech recognition system in noisy environment. In *Image and Signal Processing: 5th International Conference, ICISP 2012, Agadir, Morocco, June 28–30, 2012. Proceedings 5* (pp. 571–578). Springer.

Amrous, A. I., Debyeche, M., & Amrouche, A. (2011). Prosodic features and formant contribution for Arabic speech recognition in noisy environments. In *Advances in Intelligent and Soft Computing* (pp. 465–474).

Amrouche, A., Abed, A., & Falek, L. (2019). Arabic speech synthesis system based on HMM. In *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)* (pp. 73–78). IEEE.

Arinaitwe, P., Murungi, E., Ogenyi, F. C., Asiimwe, R., & Buhari, M. D. (2024). Review of techniques used in speech signal processing. *Deleted Journal, 3*(1), 63–70.

Boersma, P. (2006). Praat: Doing phonetics by computer (version 4.4.24). Retrieved from http://www.praat.org.

Boubakeur, K. N., Debyeche, M., Amrouche, A., & Bentrcia, Y. (2022). Prosodic modeling-based speaker identification. In *2022 2nd International Conference on New Technologies of Information and Communication (NTIC)* (pp. 1–6). Mila, Algeria.

Cui, B.-G., & Chen, X. (2010). An improved hidden Markov model for literature metadata extraction. In *Advanced Intelligent Computing Theories and Applications : 6th International Conference on Intelligent Computing, ICIC 2010, Changsha, China, August 18–21, 2010. Proceedings 6* (pp. 205–212). Springer.

Doddington, G. R. (1985). Speaker recognition—Identifying people by their voices. *Proceedings of the IEEE, 73*(11), 1651–1664.

Droua-Hamdani, G. (2020). Formant frequency analysis of MSA vowels in six Algerian regions. In *Lecture Notes in Computer Science* (pp. 128–135).

Falek, L., Amrouche, A., Fergani, L., Teffahi, H., & Djeradi, A. (2011). Formantic analysis of speech signal by wavelet transform. In *2011 Proceedings of the World Congress on Engineering, WCE 2011* (Vol. 2, pp. 1572–1576).

Fairclough, L., Brown, G., & Kirchhuebel, C. (2023). Reviewing the performance of formants for forensic voice comparison: A meta-analysis of forensic speech science research. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 3834–3838).

Ferrer, L., Scheffer, N., & Shriberg, E. (2010). A comparison of approaches for modeling prosodic features in speaker recognition. In *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 4414–4417). IEEE.

Huang, X., Acero, A., Hon, H.-W., & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice Hall PTR.

Ji, M., Wang, F., Wan, J. N., & Liu, Y. (2015). Literature review on hidden Markov model-based sequential data clustering. *Applied Mechanics and Materials, 713*, 1750–1756.

Kreiman, J., & Sidtis, D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception.*

Leu, F.-Y., & Lin, G.-L. (2017). An MFCC-based speaker identification system. In *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)* (pp. 1055–1062). Taipei, Taiwan.

Mary, L., & Yegnanarayana, B. (2006). Prosodic features for speaker verification. In *Ninth International Conference on Spoken Language Processing.*

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *International Journal of Speech Language and the Law, 13*(1), 89–126.

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of Speech Recognition.* Prentice-Hall, Inc.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3*(1), 72–83.

Singh, N., & Khan, R. (2015). Extraction and representation of prosodic features for automatic speaker recognition technology. In *Fifth International Conference on AITMC (AIM-2015), Proceedings of Advanced in Engineering and Technology* (pp. 1–7). McGraw Hill Education.

Singh, N., Khan, R., & Shree, R. (2012). MFCC and prosodic feature extraction techniques : A comparative study. *International Journal of Computer Applications, 54*(1).

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, 19–22.

Varga, A., & Steeneken, H. J. (1993). Assessment for automatic speech recognition : II. Noisex-92 : A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication, 12*(3), 247–251.

Young, S., Odell, J., et al. (2002). *The HTK Book Version 3.3.* Speech group, Engineering Department, Cambridge University Press.