DOI: 10.61782/fa.2025.0435

FORUM ACUSTICUM
EURONOISE 2025

# HUMAN-MACHINE SPEAKER IDENTIFICATION ACCORDING TO TOBI PRINCIPLES

**Bernal Ortiz, Francisco**[1*]

[1]Institute of Applied Linguistics, University of Cádiz, Spain

## ABSTRACT

Forensic linguistics has been significantly affected by advances in Generative Artificial Intelligence as there is now a wide range of tools that allow users to create spoken and written texts artificially. In addition, the accessibility of these models has caused a rise in crime rates with regard to impersonation and several other felonies. This poses a new challenge for the linguistic discipline, as speaker identification must now take into account robotic sources. Therefore, there is a need to detect those voices that are human and those produced by Large Language Models. In this pilot study, ToBI prosodic principles are studied in order to propose a feature that significantly differentiates the two types of authorship. For this purpose, a series of sentences with different prosodic characteristics have been elaborated, recordings have been made of humans and artificial intelligences, these have been annotated in a semi-automatic way to obtain their prosodic structure taking into account the mentioned principles and the similarity of these structures has been verified to obtain a conclusion.

**Keywords:** *speaker identification, ToBI, forensic linguistics, acoustic linguistics, prosody*

## 1. INTRODUCTION

In a broad conception, forensic linguistics can be defined as the interface between language and law [1]. In addition, the author explains that the object of study is the legal language, the language of legal proceedings, and the evidential language. One of the main applications of this science is speaker identification, which can be defined as "the process of extracting the identity of a speaker by using machine according to the acoustic features of the given utterance" [2, 3]. In this branch of forensic linguistics, studies have been carried out on several fields such as direct applications [4], improvement or development of computational models [2, 5] as well as overviews and literature reviews [6]. It is worth noting that this kind of identification relies entirely on acoustic phonetics which is described by [7] as:

> "The study of the physical characteristics of speech sounds as they leave their source (the speaker), move into the air, and gradually dissipate. The acoustic analysis of speech sounds requires laboratory observation with instruments and specialized (but readily available) computer hardware and software."

Tools such as Praat [8] have been developed for deep research in this regard and are of great benefit as they are considered a benefitial interface for the fields of linguistics and sound engineering as these tools provide displays of waveform and spectrogram. In Praat, there has been a recent development with the intention of systematising linguistic processes such as transcription. In that regard, the linguistics community has created Praat scripts to segment audio files in phonemes, syllables, and words, as well as automatically analysing the prosodic features of an utterance using the ToBI framework [9, 10].

*Corresponding author: paco.bernalortiz@alum.uca.es.

## 2. TOBI FRAMEWORK

Tones and Break Indices (ToBI) is a system to transcribe prosodic features of speech [11]. Specifically, it is a framework of systems which, based on the Autosegmental-Metrical model (AM), is applied to a particular language or languages [12]. In Spanish, a ToBI system was developed by [13] and has been the subject of many studies such as corpus application of the ToBI spanish system in Spanish speaking countries [14] or reviews on prosodic notation systems [15, 16]. Moreover, the spanish ToBI system has been further developed and implemented as Praat plugins by [9, 10].

As explained in [17], the ToBI system is characterised by tonal events and Break Indices. The tonal events of speech meet the following criteria:

- All accents can be low (L) or high (H). It is worth mentioning that the Spanish ToBI system also allows accents to be medium (M) as observed in [9, 10].
- Pitch accents are accents that are anchored to the stressed syllable. The star (*) in this pitch is the mark of the stressed syllable.
- Boundary tones are tones that relate to the end of an intermediate phrase or an end phrase. These are marked with a dash (-) and a percent sign (%) respectively.
- The closing exclamation mark (!) represents the small pitch whereas the opening exclamation mark (¡) represents the high pitch, and the signs less than (<) and greater than (>) denote, respectively, the delay or anticipation of the next pitch.
- A distinction is made between five Break Index groups. These have a numerical value and range from 0 as the shortest to 4 as the longest:
  - 0: the boundary of a clitic or lexical unit without a lexical accent.
  - 1: the standard boundary of a word with the following word.
  - 2: the boundary marked by a pause, but without tone marking.
  - 3: the intermediate phrase boundary.
  - 4: the boundary that ends the whole phrase.

After having developed a brief literature review, the ToBI framework has proven to be an interesting resource for studies involving other linguistic fields such as sociolinguistics [18], prosody itself [19] and clinical linguistics [20], as well as a source of further development on the branch of prosody [15, 16], a means of developing sentiment analysis [21, 22] and an incentive for automatic transcription of speech [9, 10, 23]. However, it should be noted that there is not a single research related to speaker identification on this framework. Therefore, the main objective of this pilot study is to shed light on this practice of the forensic acoustics field making use of the spanish ToBI system.

## 3. HYPOTHESIS AND OBJECTIVES

Considering the facts stated previously, the following hypotheses have been established:

- H1: It is initially believed that the tonal events labels of the sentences of robotic voices will significantly differ from those of human-generated voices.
- H2: The Break Indices labels of the robotic voices recordings will also be significantly different.

Therefore, the main objectives of this pilot study have been defined:

- O1: A test of the tonal events of every recording will be performed and a comparison of the subject groups will be carried out making use of hypothesis test.
- O2: The Break Indices of the evidence will be obtained, and the dependence of the variable will be tested using hypothesis test.

## 4. METHODOLOGY

In terms of the methodology of this study, the following steps have been taken.

### 4.1 Sentences

To maintain representativeness, a total of six sentences have been defined and separated into three pairs of sentences: enunciative, interrogative, and exclamatory. In addition, note that in each pair there is one sentence in which negation is present and one in which it is not. The sentences are as follows:

1. Enunciative sentences
   (a) La lingüística es la ciencia que estudia el lenguaje humano.
   (b) La ciudad no tiene ninguna fuente de la que beber.
2. Interrogative sentences
   (a) ¿Es verdad que en el autobús hay calefacción?
   (b) ¿No tenemos que entregar un justificante del médico?
3. Exclamatory sentences
   (a) ¡Tráeme una botella de agua de la cocina!
   (b) ¡No quiero que me digas esas cosas!

### 4.2 Study subjects

Regarding study subjects, two groups have been established. Due to the limited availability of the AI models, the robotic group has four male and four female voices. Accordingly, this is also the case for human voices. The groups are as follows:

- A sample of eight human speakers, who will be the members of the human group. This group is divided into four men and four women with Spanish from Spain as their mother tongue, between 20 and 30 years of age.
- A sample of eight Artificial Intelligence models, which will represent the robotic speech. The models are:
  - ChatGPT (version Gpt4o).
  - Gemini 2.0 Flash.
  - Google GTTS voice synthesizer (version 2.5.4).
  - Alexa (retrieved on 10/03/25).
  - Four AI models retreived from the Speechify website on 10/03/25:
    * Álvaro.
    * Enzo.
    * María.
    * Lucía.

For the sake of representativeness, the robotic subjects are of mixed natures. There are various widely known generative AIs such as ChatGPT while there are also four models of a web application. The main reason for this addition is the willingness of the subjects of the latter subgroup to imitate human speech, an idea that may not be a priority in the former subgroup.

### 4.3 Recordings

The sentences will be given to the subjects to read aloud and record, for later conversion into audio tracks. In the case of robotic voices, a recording has been made making use of the audio from the computer on which the artificial intelligences' audios have been played. The audio encoding parameters of these recordings are as follows:

- Mp3 format.
- 44100 Hz sampling rate.
- Two channels.
- 128 kb/s bitrate.

The human group recordings will be made using the recording application of a Motorola mobile phone and the Edge 40 Neo model and maintaining a distance of 15.0 cm between the speaker and the microphone. To maintain systematicity, these recordings have been converted to the other groups' format so that the audio encoding parameters remain the same for every recording.

An analysis in Praat of the audio tracks will be performed to obtain the ToBI transcription of the sentences pronounced by the different subjects. This analysis will be carried out using the Intonalyzer script [10], based on [9], which will provide a systematic method of obtaining ToBI transcription. This will return, on the one hand, the tonal events and, on the other, the Break Indices.

For Praat analysis, it is necessary first to segment the audio file in the sentence's text, its words, and phonological representation of its syllables, all in the appropriate Intonalyzer input format as seen in Fig. 1:

Once segmentation is complete, the Intonalyzer script is performed, returning results similar to Fig. 2:

The data obtained from the "BI" and "Tones" tiers will be used on the hypothesis tests. It is worth mentioning the fact that Intonalyzer omits level 3 Break Indices by design; however, as the process after the initial segmentation is supposed to be automatic, a manual readjustment will not be carried out.

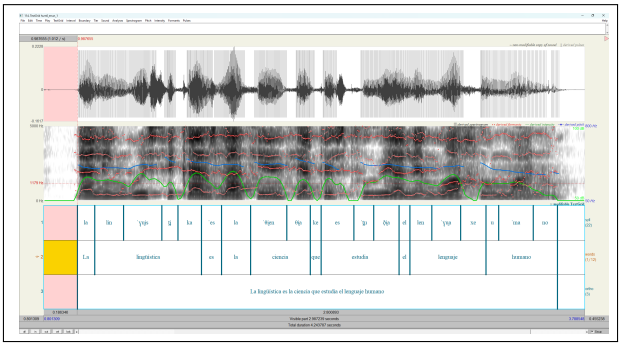

**Figure 1**. Segmented sentence example.

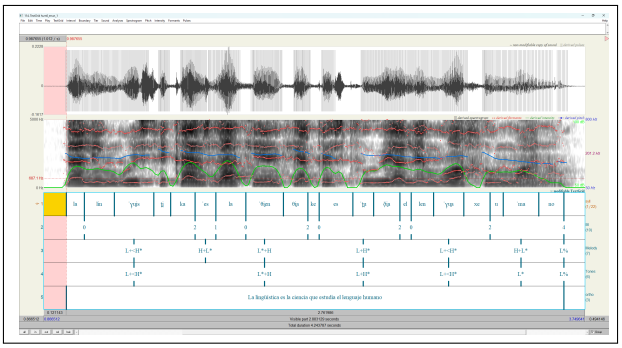

**Figure 2**. Intonalyzer output example.

### 4.4 Hypothesis testing

Using hypothesis testing, it will be assessed whether there are significant differences between the two groups of speakers. To do this, the data will be segmented into several groups:

1. Results of all values per speaker group.
2. Results of tonal events per speaker group.
3. Results of Break Indices per speaker group.
4. Results of all values per speaker group and type of sentence.
5. Results of tonal events per speaker group and type of sentence.
6. Type of Break Index per speaker group and type of sentence.
7. Type of all values per speaker group and occurrence or absence of negation.
8. Type of tonal events per speaker group and occurrence or absence of negation.
9. Type of Break Index per speaker group and occurrence or absence of negation.

As the data handled consist of categorical variables, a contingency matrix of each set of results will be elaborated, and the chi-square hypothesis test will be performed. In the case of this study, the significance level of the test is established to 0.05.

Will all of the above, the main conclusions of the study will be drawn, accepting or rejecting the initial hypothesis, and future research issues will be assessed.

## 5. RESULTS

### 5.1 Frequency tables of ToBI transcriptions

The data for the different ToBi transcriptions are the following:

**Table 1**. Break Indices per occurrence or absence of negation

| Group | Negation | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| Robotic | No | 79 | 19 | 83 | 0 | 27 |
| | Yes | 64 | 24 | 86 | 0 | 28 |
| Human | No | 80 | 20 | 81 | 0 | 27 |
| | Yes | 64 | 27 | 84 | 0 | 25 |

**Table 2**. Break Indices per type of utterance

| Group | Sentence | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| Robotic | Enun. | 63 | 9 | 67 | 0 | 21 |
| | Excl. | 40 | 16 | 48 | 0 | 16 |
| | Inte. | 40 | 16 | 54 | 0 | 18 |
| Human | Enun. | 64 | 18 | 59 | 0 | 18 |
| | Excl. | 40 | 12 | 52 | 0 | 16 |
| | Inte. | 40 | 17 | 54 | 0 | 17 |

As shown in Tab. 1 and Tab. 2, there is little variability among the data obtained regarding Break Indices. Nevertheless, it may be worth mentioning the difference of the level 1 Break Index in enunciative sentences in Tab. 2, where the human group doubles the robotic one. Nonetheless, the observed results seem to be so homogeneous as to reject apparent differences in this regard.

**Table 3**. Type of tonal event per occurrence or absence of negation.

| Group | Negation | H% | H* | H*+L | H+L* | HH% | HL% | L% | L* | L*+H | L+¡H* | L+H* | L+H*+L | L+¡H* | L+¡H*+L | LH% | LM% | M% | ¡H* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Robotic | No | 0 | 3 | 2 | 3 | 7 | 0 | 19 | 13 | 2 | 49 | 25 | 6 | 1 | 2 | 0 | 0 | 1 | 0 |
| | Yes | 2 | 5 | 0 | 0 | 7 | 2 | 12 | 17 | 4 | 37 | 42 | 4 | 2 | 0 | 1 | 1 | 1 | 0 |
| Human | No | 1 | 4 | 1 | 9 | 4 | 2 | 14 | 24 | 10 | 34 | 15 | 2 | 4 | 0 | 0 | 2 | 3 | 1 |
| | Yes | 2 | 3 | 3 | 5 | 4 | 2 | 14 | 28 | 9 | 37 | 18 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |

**Table 4**. Type of tonal event per type of utterance.

| Group | Sentence | H% | H* | H*+L | H+L* | HH% | HL% | L% | L* | L*+H | L+¡H* | L+H* | L+H*+L | L+¡H* | L+¡H*+L | LH% | LM% | M% | ¡H* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Robotic | Enun. | 2 | 2 | 0 | 1 | 1 | 1 | 15 | 14 | 2 | 26 | 36 | 3 | 2 | 0 | 1 | 0 | 1 | 0 |
| | Excl. | 0 | 2 | 1 | 0 | 0 | 0 | 15 | 8 | 1 | 24 | 19 | 6 | 1 | 2 | 0 | 1 | 0 | 0 |
| | Inte. | 0 | 4 | 1 | 2 | 13 | 1 | 1 | 8 | 3 | 36 | 12 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| Human | Enun. | 2 | 0 | 1 | 7 | 2 | 0 | 14 | 13 | 8 | 30 | 15 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| | Excl. | 0 | 2 | 1 | 6 | 0 | 2 | 11 | 17 | 6 | 22 | 8 | 1 | 3 | 0 | 0 | 2 | 0 | 0 |
| | Inte. | 1 | 5 | 2 | 1 | 6 | 2 | 3 | 22 | 5 | 19 | 10 | 2 | 1 | 0 | 1 | 1 | 3 | 1 |

In the case of tonal events, there are clear differences between the distribution of tonal events in the groups. For instance, the interrogative sentences articulated by humans show three times more low-pitch accents (L*), as well as a clear avoidance of robotic voices to perform a binomial tonal event starting high and ending in a low-pitch accent (H+L*) in the case of both sentences with and without negation, with a frequency of 0 in the case of robotic sentences with occurrence of negation.

Nevertheless, there are also various tonal events that show similar distributions, such as the binomial tonal event starting low and following with a small high-pitch accent (L+!H*).

### 5.2 Chi-square tests results

After carrying out the hypothesis tests mentioned above in 4.4 between the robotic and human voices groups, the results in Tab. 5 show several points of interest.

Firstly, every chi-square test that involves tonal events has a significantly low p-value. This value is even lower when this is the only feature, reaching a $1.24E-09$ in the case of discerning tonal events grouping by subject groups and type of utterance or $2.5E-04$ in the case of tonal events by subject group.

However, this is not the case for the Break Indices, which seem to show little variability between the subject groups and a lack of significant difference in the tests where this variable is the only one included. Moreover, this feature seems to affect negatively the results of the tests where these features are taken into account, returning p-values close to 1 on the tests that this is the only feature.

## 6. DISCUSSION

It is worth mentioning the fact that the manual classification of the generative AIs robotic subgroup does not pose a challenge by any means and was clear after an initial hearing of the recordings. The reason may be that they do not present variability in the intonational curve of the different sentences and a small amount of the recordings are somewhat difficult to understand. For this reason, it seems that the voice synthesis models of these AIs in Spanish do not aim to clone human voices, or at least, they do not succeed in this task. In contrast, they seem to aim to perform a simple voice synthesis of the text provided. In spite of that, the other subgroup of robotic voices does indeed pose a challenge to the human ear.

Additionally, there has been an unexpected lack of availability of robotic voices in Spanish of Spain to carry out the investigation. The main inconvenient is the intrinsic limitation of the study in recording the exact same utterances in all cases. Taking that into account, it would be advisable to overcome this limitation for further research.

Furthermore, there are several matters to take into account for future work:

- More informants would be required to perform a more representative study of speaker identification. In addition, a wider range of robotic voices is preferred to ensure representativeness in this regard.
- The repeating combination of tonal events should be evaluated and tested. Such values may indicate a speech systematicity on any of the groups of speakers as well as significant differences between the two.

**Table 5**. Chi-square results

| Test | $\chi^2$ | *pvalue* | *df* |
|---|---|---|---|
| **All values by group** | **45.31** | **1.6E-03** | **21** |
| **Tones by group** | **44.93** | **2.5E-04** | **17** |
| BI by group | 0.31 | 0.95 | 3 |
| **All values by group and utterance** | **199.48** | **7.65E-08** | **105** |
| **Tones by group and utterance** | **187.03** | **1.24E-09** | **85** |
| BI by group and utterance | 11.04 | 0.75 | 15 |
| **All values by group and negation** | **84.18** | **3.9E-02** | **63** |
| **Tones by group and negation** | **78.65** | **7.7E-03** | **51** |
| BI by group and negation | 5.27 | 0.81 | 9 |

- It would be highly beneficial to replicate the study using a more segmented prosodic analysis, with the counterpart of, probably, researching and creating such kind of analysis.
- It may also be advisable to implement a system that studies prosody more demandingly, as well as extracting repeating patterns or similarities regarding aspects such as the intonation curve.
- As for further research, it would be highly beneficial to create a logistic regression or a model of the likes to be able to make predictions rather than simply performing hypothesis tests and observing differences.

## 7. CONCLUSIONS

As seen in the results, there are significant differences between humans and machines in the analysis of ToBI annotation of the proposed sentences. The chi-square tests indicate tonal events to be the most discriminatory feature, whereas the Break Indices do not behave likewise.

On the one hand, it is safe to consider tonal events as significantly different for human and robotic voices in Spanish of Spain using the Spanish ToBI system provided by [9,10], accepting the first established hypothesis. These results imply that the tone is worth further consideration and research.

On the other hand, Break Indices do not seem to present significant variability between the studied groups, as the p-values are nowhere near the established significance level of 0.05. It is without a doubt a variable to exclude in the case of speaker identification, and thus the second hypothesis is rejected.

To sum up, the ToBI framework seems to be an effective and interesting starting point to further improve speaker identification in Spanish, always taking into consideration that the results may vary with a greater number of subjects. However, it should be noted that the Break Indices are a feature to avoid.

## 8. REFERENCES

[1] M. Ramírez Salado, "Imprecisiones terminológicas derivadas de la traducción en el ámbito de la lingüística forense," *Revista de Lingüística y Lenguas Aplicadas*, vol. 16, pp. 175–183, July 2021.

[2] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi, and I. Ali, "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, p. 114591, 2021.

[3] G. Humblot-Renaux, C. Li, and D. Chrysostomou, "Why talk to people when you can talk to robots? far-field speaker identification in the wild," in *Proc. of the 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 272–278, IEEE, 2021.

[4] J. Campbell, W. Shen, W. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[5] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Applied Sciences*, vol. 11, no. 8, 2021.

[6] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[7] M. Umiyati, "A literature review of forensic linguistics," in *Proc. of the International Journal of Forensic Linguistics*, 2020.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.4.27)," 2025.

[9] W. Elvira-García, P. Roseano, A. M. Fernández-Planas, and E. Martínez-Celdrán, "A tool for automatic transcription of intonation: Eti_tobi a ToBI transcriber for Spanish and Catalan," *Language Resources and Evaluation*, vol. 50, no. 4, pp. 767–792, 2016.

[10] J. M. Lahoz-Bengoechea, "Intonalyzer: A semi-automatic tool for Spanish intonation analysis (1.0)," 2021. original-date: 2021-05-29T19:36:56Z.

[11] University of Ohio, "ToBI," 1999.

[12] J. B. Pierrehumbert, *The phonology and phonetics of English intonation*. Thesis, Massachusetts Institute of Technology, 1980. Accepted: 2009-01-23T14:36:47Z.

[13] M. E. Beckman, M. Díaz-Campos, J. T. McGory, and T. A. Morgan, "Intonation across Spanish, in the Tones and Break Indices framework," *Probus*, vol. 14, Jan. 2002.

[14] H. Ortiz Lira, "La aplicación de ToBI a un corpus del español de Chile," *Onomázein*, pp. 429–442, Dec. 1999.

[15] E. Estebas Vilaplana and P. Prieto Vives, "La notación prosódica del español: una revisión del Sp_tobi," *Journal of Experimental Phonetics*, vol. 17, pp. 263–283, Dec. 2008. Section: Articles.

[16] Sosa, J. M., "La notación tonal del español en el modelo SP_tobi.," in *Teorías de la Entonación*, pp. 185–208, Barcelona: Ariel Lingüística, prieto, p. ed., 2003.

[17] S.-A. Jun, "The ToBI Transcription System: Conventions, Strengths, and Challenges," in *Prosodic Theory and Practice* (J. Barnes and S. Shattuck-Hufnagel, eds.), p. 0, The MIT Press, Feb. 2022.

[18] N. R. Holliday, "Perception in Black and White: Effects of Intonational Variables and Filtering Conditions on Sociolinguistic Judgments With Implications for ASR," *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[19] J. F. Pitrelli, "ToBI prosodic analysis of a professional speaker of American English," in *Proc. of Speech Prosody 2004*, pp. 557–560, ISCA, Mar. 2004.

[20] F. Chen, C. C.-H. Cheung, and G. Peng, "Linguistic Tone and Non-Linguistic Pitch Imitation in Children with Autism Spectrum Disorders: A Cross-Linguistic Investigation," *Journal of Autism and Developmental Disorders*, vol. 52, pp. 2325–2343, May 2022.

[21] L. Shen and W. Wang, "Improving Speech Emotion Recognition Based on ToBI Phonological Representations," *PATTERNS 2018: the Tenth International Conference on Pervasive Patterns and Applica- tions.*, pp. 1–5, 2018.

[22] A. I. Iliev, Y. Zhang, and M. S. Scordilis, "Spoken Emotion Classification Using ToBI Features and GMM," in *Proc. of the 2007 14th International Workshop on Systems, Signals and Image Processing and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services*, (Maribor, Slovenia), pp. 495–498, IEEE, June 2007.

[23] W. Zhai and M. Hasegawa-Johnson, "Wav2ToBI: a new approach to automatic ToBI transcription," in *Proc. of INTERSPEECH 2023*, pp. 2748–2752, ISCA, Aug. 2023.