Was That a Question?
Automatic Classification of Discourse Meaning in Spanish

Santiago Arróniz
Indiana University
sarroniz@indiana.edu

Sandra Kübler
Indiana University
skuebler@indiana.edu

Abstract

This paper examines the effectiveness of different feature representations of audio data in accurately classifying discourse meaning in Spanish. The task involves determining whether an utterance is a declarative sentence, an interrogative, an imperative, etc. We explore how pitch contour can be represented for a discourse-meaning classification task, employing three different audio features: MFCCs, Mel-scale spectrograms, and chromagrams. We also determine if utilizing means is more effective in representing the speech signal, given the large number of coefficients produced during the feature extraction process. Finally, we evaluate whether these feature representation techniques are sensitive to speaker information. Our results show that a recurrent neural network architecture in conjunction with all three feature sets yields the best results for the task.

1 Introduction

The aim of this study is to investigate the efficacy of feature representations of audio data in accurately classifying discourse meaning in Spanish. The task involves determining whether an utterance is a declarative sentence, an interrogative, an imperative, etc. Since there does not seem to be an agreed upon name for this task, we will refer to it as discourse meaning (rather than referring to a broader sense of this term).

In human perception, this process involves the comprehension of the relationship of words, phrases, and clauses used in a sentence, as well as their overall contribution to the intended meaning of the sentence. We focus on the prosodic features of different discourse meanings in Spanish. Pitch, or the perceived highness or lowness of a sound, can play a role in distinguishing between different discourse meanings. For example, declarative sentences typically present a falling pitch contour, indicating that the statement is complete, while interrogatives usually have a rising contour, signaling that a question is being asked.

In contrast to tonal languages such as Mandarin Chinese, Thai, or Punjabi, which mark the phonological contrast of pitch at the lexical level (word level), intonational languages such as Spanish or English mark the phonological contrast of pitch at the utterance level. For Spanish, pitch movements are mainly used to signal discourse meaning or to mark focus. The properties that govern production in intonation are structurally analogous to those that govern lexical tones and morphological paradigms (Ladd, 2008). This means that a declarative statement like *María viene* 'María is coming' and its interrogative counterpart *¿María viene?* 'Is Maria coming?' differ only in the intonational contour with which they are produced, since both are syntactically and lexically identical.

1.1 Research Questions

Our study focuses on three main research questions:

RQ1: How do we represent intonation as features in discourse-meaning classification for Spanish?

RQ2: Do different feature representations convey distinct types of information?

RQ3: Are these feature representation methods sensitive to speaker information, or do they abstract away from this information?

RQ1 addresses the question of how pitch contour information can be represented for a discourse-meaning classification task using speech data of Spanish. We focus on three different audio features widely used in speech recognition and classification tasks such as emotion recognition (Badr et al., 2021; Issa et al., 2020; Zhou et al., 2019): Mel Frequency Cepstral Coefficients (MFCCs), Mel-scale

Proceedings of Recent Advances in Natural Language Processing, pages 132–142
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_015

spectrograms, and chromagrams. We also evaluate the effectiveness of using mean values of each band as opposed to all frequency measures. Utilizing means may be efficient when representing the speech signal for a discourse-meaning classification task, given the large number of coefficients produced during the feature extraction process.

RQ2 is concerned with the differences between the three audio feature representations. If they convey different types of information, we expect to see improvements in classification by using combinations of representations. MFCCs, generally considered one of the most effective type of feature in audio classification tasks (Dave, 2013; Xie and Liu, 2006), discard a significant amount of information by a low-rank linear projection of the Mel spectrum. Thus, Mel spectrograms and chromagrams may provide information that is no longer present in MFCCs.

RQ3 examines potential speaker effects in our data. Specifically, we investigate if there are individual differences in how people produce the intonation curves for distinct discourse meanings, and whether the feature representations are sensitive to those differences; i.e., whether these audio representations can generalize across different discourse meanings, or if there is any overlap that could lead to bias in the classification process.

The remainder of the paper is organized as follows: Section 2 outlines previous research on Spanish intonation and modeling intonation in other languages. Section 3 details the methodology utilized in this study, including information about the corpus, the algorithms, the feature extraction processes, and hyperparameter optimization. Sections 4 – 6 present the results for the three research questions. Finally, Section 7 outlines our conclusions and future work.

2 Related Work

2.1 Spanish Intonation

Spanish sociophonetic research (Face, 2001, 2005, 2008, 2004; Estebas-Vilaplana and Prieto, 2010; Quilis, 1993) describes the pitch contours used by speakers in different dialectal areas. The majority of intonation studies conducted in Spain are descriptive, with a focus on describing the intonational contours of certain regions. Many of these studies have relied on elicited speech to analyze these productions (e.g., Estebas-Vilaplana and Prieto, 2010), while others have adopted a corpus approach (e.g., Torreira and Floyd, 2012). However, it remains to be explored how generalizable these contours are, and whether machine learning techniques can be applied to extract information about intonation and automatically classify discourse meaning.

2.2 Speech Classification

The automatic detection and classification of discourse meaning has been the focus of many recent studies in speech classification. Prosody modeling has been particularly important in English and other languages, with research focused on detecting prominence and phrase boundaries (Levow, 2005). Researchers have explored incorporating context into feature-level recognition of prosodic events (Mishra et al., 2012), as well as normalizing features by immediate context when detecting and classifying prosodic events (Rosenberg, 2009, 2010, 2012). Sequential models have also been used to examine prosodic modeling, with some studies attempting to predict prominence and phrasing at the syllable and word level using models based on normalized segment duration and pauses (Wightman and Ostendorf, 1994; Ananthakrishnan and Narayanan, 2005).

Additionally, modeling F0 contours has been explored; some of them attempted to model F0 contours directly (Bailly and Holm, 2005; Fujisaki, 1983; Hirst and Espesser, 1993; Kochanski and Shih, 2003; Ni et al., 2006; Pierrehumbert, 1981; Taylor, 2000; Van Santen and Möbius, 2000), while others simulated the underlying mechanisms of F0 production (Chodroff and Cole, 2019; Cole et al., 2022). Most recent studies have used deep learning models such as LSTM neural networks (Zeyer et al., 2017; Sundermeyer et al., 2012), and multimodal deep learning approaches that combine audio and text inputs to achieve high performance on speech intention classification tasks (Gu et al., 2017; Agüero and Bonafonte, 2004). However, more research is needed to explore how different machine learning approaches can be used to model intonation in languages such as Spanish.

3 Methodology

3.1 Corpus

For our experiments, we collected a scripted speech corpus[1] that was designed for the analysis of Span-

[1] https://github.com/sarroniz/speech_corpus

ish intonation under laboratory conditions, to exclude factors such as the length of utterances, differences in lexical content, noise in the signal, etc. The reading task included six different types of discourse meaning, each having a total of twenty examples. The elicited discourse meanings (Hualde and Prieto, 2015) are described below, the corresponding schematic representation of the contours are shown in Figure 1.

Broad Focus Declarative Statements are the most common type of discourse meaning. They are used to bring every element in the sentence into focus, so there is no emphatic element in the utterance (e.g., *Juan compra pan* 'Juan buys bread'). The syntactic structure in Spanish is usually subject (S), verb (V), and complements (C).

Narrow Focus Statements selectively focus on one part of the sentence (e.g.: *Juan compra pan* 'Juan buys bread' as the answer to the question *¿Quién compra pan?* 'Who buys bread?', where *Juan* is focused information). The syntactic structure is usually SVC.

Absolute Interrogatives are used to request a yes/no answer from the interlocutor. Spanish yes/no questions have the same syntax as broad focus statements, and require intonation to convey interrogativity in the absence of contextual cues. Unmarked questions may omit the inversion of the subject, but it is often omitted (e.g., *Compran pan* 'They buy bread' vs. *¿Compran pan?* 'Do they buy bread?').

Partial Interrogatives are interrogative sentences that convey interrogativity directly through the presence of a question word, without the need for intonational signaling (e.g.: *¿Quién viene a la fiesta?* 'Who is coming to the party?'). Unmarked partial interrogatives in Spanish can share the same intonation pattern as broad focus statements.

Exclamatives are utterances with an exclamative nuance and show an initial peak in the nuclear accent that aligns within the accented syllable (e.g.; *¡Qué mañana tan bonita!* 'What a lovely morning!').

Imperatives in Spanish are often highly exclamatory, resulting in an expanded pitch range, greater intensity, and longer duration. Their intonation patterns can vary and are not necessarily linked to specific geographic regions. Imperatives are often represented by final pitch accents.

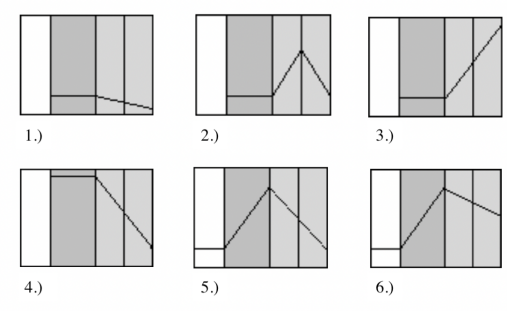

Figure 1: Schematic representations of the contours in Spanish for the six types of discourse meaning (Estebas-Vilaplana and Prieto, 2010).

We collected samples from nine different speakers (seven from southern Spain, and two from the Madrid area). In total, 1 080 different speech productions (9 speakers * 6 types of discourse meanings * 20 examples) were used for our experiments, with an average duration of 1.159 seconds[2]. For all of the audios, the corpus includes information about demographic information of the speakers (such as age, gender, level of education, time spent out of their place of birth, etc.), plus the type of discourse meaning.

3.2 Classifiers

We experiment with different classifiers using the *scikit-learn* library (Pedregosa et al., 2011): support vector machines (SVC), Random Forest, k-nearest-neighbors (kNN), decision trees, and a multilayer Perceptron (MLP). We use grid search cross-validation to optimize hyperparameters.

Additionally, we experiment with Long Short-Term Memory (LSTM) recurrent neural networks, both unidirectional and bidirectional, using *Keras* in TensorFlow[3]. The model takes input in the form of a 1-dimensional sequence, where the length of the sequence is determined by the number of features in the input data. Three convolutional layers are stacked; each layer consists of a convolutional operation followed by batch normalization, activation (using the ELU activation function), max pooling, and dropout. The LSTM layer was added with 64 units. We set the model to return sequences rather than just the last output. We also use a softmax activation function in a fully connected dense layer. We follow the hyperparameter optimization by Zeyer et al. (2017) for acoustic modeling in

[2]Only sonorant segments were included (no occlusives), resulting in a continuous, uninterrupted pitch signal.

[3]tensorflow.org

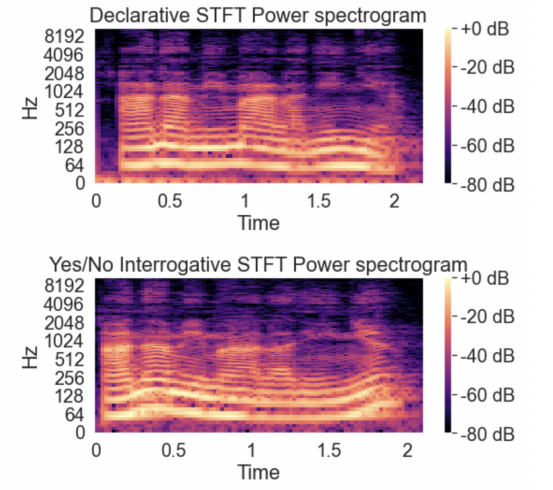

Figure 2: STFT Spectrogram examples for a declarative sentence (top) vs. an absolute yes/no interrogative sentence (bottom).

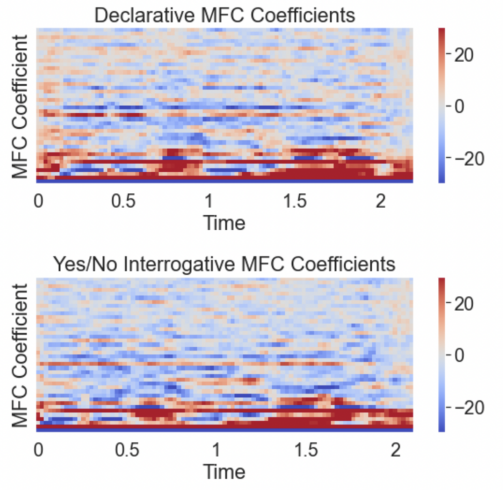

Figure 3: MFCC representations for a broad focus declarative sentence (top), and an absolute yes/no interrogative sentence (bottom).

speech recognition.

For the optimal hyperparameters used in the experiments, see the Tables in appendix A.

3.3 Feature Extraction

We use three different audio feature representations: *Mel-Frequency Cepstral Coefficients (MFCCs)*, *Mel spectrograms*, and *chromagrams*. MFCCs are commonly used in speech recognition systems (Dave, 2013) and represent the spectral envelope of speech, while Mel spectrograms are a spectral representation of audio signals where the frequency scale is warped to better match human auditory perception. Chromagrams, in contrast, are a type of harmonic feature that capture the pitch content of an audio signal by projecting the frequency content onto a set of pitch classes.

All three feature sets are extracted using *librosa* (McFee et al., 2015), a Python library for audio analysis and feature extraction. We start by extracting the Short-Time Fourier Transform (STFT) of each audio sample. By computing the Fourier transform on each segment, multiple power spectrograms are produced for each audio file. The frame size and hop size are set to default in librosa ('n_fft=2048' and 'hop_length=512'). Figure 2 shows two examples of the STFT power spectrograms.

Mel-Frequency Cepstral Coefficients We use triangular, overlapping window functions (Hanning function) on the STFT power spectra and compute the energy within each window. Then we map the frequencies to the Mel scale. After testing a range of coefficients for MFCCs (10, 20, 40, and 60), we choose 40 since it proved optimal during optimization. Figure 3 shows two examples of the MFC coefficients representations. Positive MFCCs correspond to low-frequency regions of the cepstrum, and negative MFCCs represent high-frequency regions.

Mel Spectrogram Mel spectrograms convert the frequency axis of a spectrogram to a non-linear Mel scale, which is based on the human auditory system's response to frequency[4]. Mel frequencies are logarithmically spaced, and equal distances on the Mel scale correspond to equal perceptual differences in pitch. We generate Mel spectrograms using a filterbank of triangular overlapping filters that sum to 1 over the frequency axis of the spectrogram. The resulting coefficients represent the energy in a particular Mel frequency bin at a specific time. Figure 4 shows two examples of Mel spectrograms.

Chromagrams provide a mapping of the audio signal to pitch classes over time, i.e.; CDEFGAB plus five semitones (Birajdar and Patil, 2020). Chromagrams are computed by grouping the STFT coefficients into 12 frequency bands, resulting in a 12-dimensional feature vector for each time frame. Figure 5 shows two examples of chromagrams.

[4]Mel spectrograms are similar to MFCCs, the difference stems from the use of a nonlinear Mel-scale frequency axis instead of the linear frequency axis of traditional spectrograms.

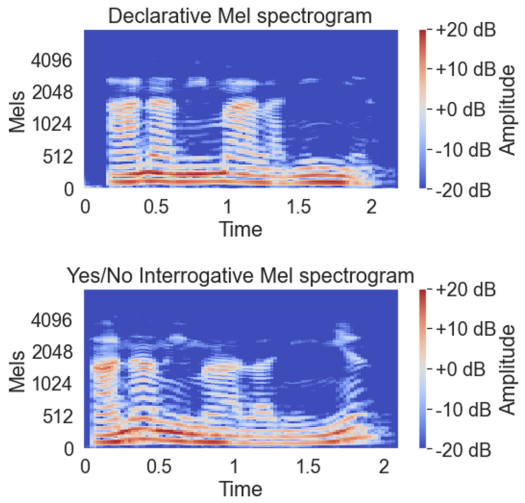

Figure 4: MEL frequency spectrograms, for a broad focus declarative sentence (top) and an absolute yes/no interrogative sentence (bottom).

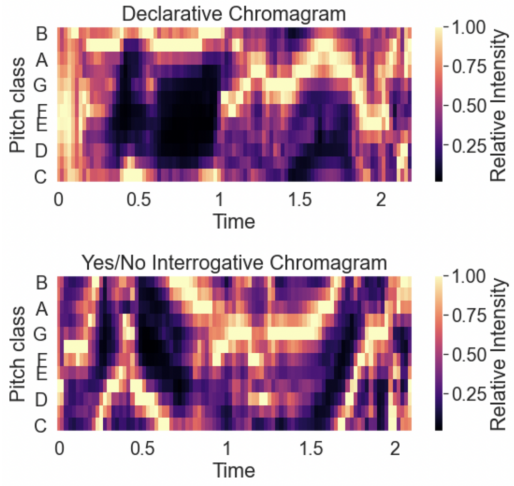

Figure 5: Chromagram examples for a broad focus declarative sentence (top) and for an absolute yes/no interrogative sentence (bottom).

3.4 Data Normalization and Scaling

After creating matrices of the three feature sets under consideration, we scale the resulting features, standardizing the different coefficients so that they have zero mean and unit variance (using *StandardScaler* in scikit-learn).

4 RQ 1: Exploring Audio Feature Representations

The first research question (RQ1) investigates the effectiveness of the three representations of the audio signal: MFCCs, Mel-scale spectrograms, and chromagrams. Our goal is twofold: 1) to investigate whether the three audio features are effective in capturing the necessary information to classify pitch based on discourse meaning, and 2) to assess whether the use of mean values, as opposed to all values, is a more efficient method for capturing this information.

Since the number of frames produced by STFT varies based on the length of each audio file, the exact number of all the coefficients for each feature set varied accordingly. Therefore, to ensure uniformity, we padded with zero values such that each file had the same number of coefficients as the longest file. Specifically, we set the number of coefficients to MFCCs=7,840; Mel spectrograms=23,936; chromagrams=3,812 (see Table 1). In the case of means, we generated a matrix back from each extraction process, and computed the mean of those matrices to obtain a single feature array for each speech sample. We obtained a total of 180 features for each array, distributed as follows (see Table 1): MFCCs=40; Mel Spectrograms=128 (number of Mel frequency bands); chromagrams=12 (one per pitch class).

Features	All values (N)	Means (N)
MFCC	7 840	40
Mel Spectrogram	23 936	128
Chromagram	3 812	12

Table 1: Distribution of the number of coefficients for each feature set for each audio sample when using a) all the values provided by STFT, and b) the means of those values for all the frames along the time axis.

We performed a stratified, randomized 9-fold cross-validation for each of the experiments in order to compare these results with those for RQ3 below (we had 9 speakers in our corpus).

4.1 Results and Discussion

The results for RQ1 are shown in Table 2. Overall, we see that the performance of the algorithms varies significantly depending on the feature type used. Among the algorithms tested, the LSTM models perform the best when using the feature types Mel spectrograms and chromagrams while the MLP outperforms both LSTM models when using MFCC, reaching the highest accuracy of 82.64% (using means).

In terms of feature types, MFCCs and Mel spectrograms outperform chromagram features across all classifiers. MFCCs yield the highest accuracy for every algorithm (ranging from 35.08% to

Classifier	MFCC		Mel		Chrom	
	means	all	means	all	means	all
Random Forest	58.91	36.65	56.95	27.50	42.29	39.06
KNN	69.79	37.75	63.54	22.13	47.01	44.53
SVC (linear)	68.55	38.53	54.49	21.86	39.65	41.80
SVC (RBF kernel)	53.34	37.31	50.35	21.91	43.14	40.36
Decision Tree	53.32	35.08	51.89	23.40	41.15	37.47
MLP	**82.64**	40.10	59.13	34.90	42.25	57.92
LSTM	79.16	54.17	63.14	37.50	50.64	29.17
BiLSTM	80.55	45.83	**68.33**	45.83	**54.72**	41.67

Table 2: Results for the different combination of audio representations for each model.

82.64%), followed by Mel spectrograms (ranging from 21.86% to 68.33%), whereas chromagram features yield the lowest accuracy (ranging from 29.17% to 54.72%). When considering all the features, the LSTM model with MFCCs using means achieved the highest accuracy (82.84%), while the LSTM model with chromagram features using means result in the lowest accuracy (29.17%).

In general, the use of means provides better results than using the individual values extracted from STFT frames, with around 20-30% of improvement in most cases. The only exceptions are the linear SVC and MLP used with chromagrams, which see a slight decrease in their accuracy when using means instead of all the coefficients (e.g., from 41.80 to 39.65 for the linear SVC).

Using the means of the values in MFCCs can be beneficial because it reduces the dimensionality of the feature set, making it less prone to overfitting and noise. Using mean values captures essential information in the audio signal while avoiding noise and irrelevant variations in individual frames. Mean values also provide more global information about the signal. For chromagrams, this approach may be more effective due to their high dimensionality and the need to capture harmonic and inharmonic relationships between musical notes, while also mitigating overfitting and computational complexity issues.

The findings of this experiment indicate that employing means of MFCC features in combination with an MLP yields the most effective classifier for the precise categorization of discourse meaning in Spanish. However, further investigation is required to understand the specific information conveyed by each feature, and whether combining them will lead to an improvement in classification performance.

5 RQ2: Comparing Information Content of Audio Features

RQ2 investigates whether the three audio feature representations convey different types of information. While MFCCs have been shown to be the most effective for the audio classification tasks for RQ1, their reliance on a low-rank linear projection of the Mel spectrum may lead to information loss. Thus, we explore the possibility of enhancing the discriminatory power of MFCCs by incorporating additional representations, such as Mel spectrograms or chromagrams, which may convey complementary information. If the combinations of audio features provides a full set of information, we expect increased classification results.

We focus on means for each feature type since their use resulted in higher accuracy for RQ1. We replicate the methodology of the previous experiment using the same data split as above.

5.1 Results

Table 3 shows the results from this experiment (for ease of comparison, we repeat the 'means' results from Table 2). Overall, the results show that the combination of features has a significant impact on classification accuracy, either positive or negative: When combining MFCCs and Mel spectrograms, all classifiers profit from the addition of Mel spectrograms in comparison to using only MFCCs. In this setting, MLP reaches the highest accuracy of 83.80%. In contrast, adding chromagrams to MFCCs results in a decrease in accuracy for all models, except for the LSTMs, which show an increase in accuracy from 80.55% (MFCC) to 81.94% for the combined-features model (for the biLSTM). However, this is still minimally lower than the MLP's results using this combination of features.

Classifier	MFCC	Mel	Chrom	MFCC+Mel	MFCC+Chrom	Mel+Chrom	All
Random Forest	58.91	56.95	42.29	60.99	58.57	58.80	60.34
KNN	69.79	63.54	47.01	70.70	66.67	66.93	71.70
SVC (linear)	68.55	54.49	39.65	70.77	65.88	58.72	70.90
SVC (RBF)	53.34	50.35	43.14	54.95	51.51	53.30	55.03
Decision Tree	53.32	51.89	41.15	55.05	52.31	53.35	53.97
MLP	**82.64**	59.13	42.25	**83.80**	**82.18**	65.05	**84.61**
LSTM	79.16	63.14	50.64	82.86	79.62	66.20	83.14
BiLSTM	80.55	**68.33**	**54.72**	81.75	81.94	**68.89**	83.05

Table 3: Results for the different combination of audio features per classifier.

When we combine Mel spectrograms with chromagrams, we observe a slight increase in accuracy of around 3-5% for most classifiers over the performance of the individual models. However, even the best model (using the biLSTM, reaching 68.89%) is about 11.5% lower than when combining the biLSTM with MFCCs (80.55%).

The performance of the combination of all three feature types is generally very close to that of the MFCC+Mel combination, thus showing that chromagrams do not add much additional information to the mix. Most classifiers profit minimally from the addition of chromagrams. The only exceptions are the random forest, and the decision tree. The biLSTM reaches the highest performance overall with an accuracy of 84.68%.

The results from this experiment indicate that combining MFCCs and/or Mel spectrograms with chromagram features can enhance the accuracy of our classification tasks. Chromagrams capture distinct information from MFCC and Mel spectrograms, and while they do not have enough discriminative power on their own, they introduce some new information to the other features. However, not all classifiers can profit from the addition of information, we see an intricate interaction of classifier type, feature type, and performance.

6 RQ3: Analyzing Speaker Effects

RQ3 investigates the impact of speaker effects on the classification of discourse meaning. Our objective is to examine whether there exist individual variations in how people generate intonation curves for different types of sentences and whether these differences are captured by the three feature representations.

We replicate the previous experiments while employing a leave-one-out cross-validation approach where each fold corresponds to one speaker. Since the model has not seen any data from the test speaker, a deterioration in this setting will indicate that the features types include speaker specific information.

6.1 Results and Discussion

Results for the experiment with individual features are shown in Table 4, while Table 5 shows the results for the combination of features. Columns labeled 'Random' show the results from RQ1 and RQ2 for reference, and columns labeled 'Speaker' show the results when we split by speaker.

The results in Table 4 show the expected pattern, the results when leaving out a speaker are generally lower than the corresponding random settings. The only exception is for the MLP using Mel spectrograms, for which the results improve marginally (from 59.13% to 59.26%). The smallest decreases occur when using the MFCC and non-neural methods. The results of the LSTMs decrease by more than 10% absolute with all feature types, even when using MFCCs. For the Mel spectrograms and chromagrams, these losses are more similar to those of the non-neural classifiers, which also suffer significant losses. The highest results are once again obtained when using the MLP with MFCCs, reaching 81.13%, which is only slightly lower than the 82.64% in the corresponding random setting.

The results for the combination of features in Table 5 show the same trend: Splitting the data by speaker causes slight to significant losses across the different classifiers and feature combinations. The same combinations that work well for the random data split also work well for the speaker setting. We obtain the best results using the MLP with all features (84.49%).

Overall, these results show that, as expected, there is speaker dependent information present in the features. If we do not have access to an example

	MFCC		Mel Spectrogram		Chromagram	
Classifier	Random	Speaker	Random	Speaker	Random	Speaker
Random Forest	58.91	57.59	56.95	49.17	42.29	36.94
KNN	69.79	65.28	63.54	54.86	47.01	39.81
SVC	68.55	64.29	54.49	44.56	39.65	34.49
SVC (RBF kernel)	53.34	52.16	50.35	39.20	43.14	37.73
Decision Tree	53.32	52.14	51.89	42.82	41.15	36.15
MLP	82.64	81.13	59.13	59.26	42.25	41.78
LSTM	79.16	67.59	63.14	53.98	50.64	24.62
BiLSTM	80.55	68.14	68.33	52.12	54.72	42.96

Table 4: Results comparing random data splitting to leaving out an individual speaker.

	MFCC+Mel		MFCC+Chrom		Mel+Chrom		all	
Classifier	Random	Speaker	Random	Speaker	Random	Speaker	Random	Speaker
Random Forest	60.99	58.22	58.57	56.87	58.80	50.12	60.34	57.11
KNN	70.70	65.51	66.67	65.28	66.93	57.52	71.70	65.51
SVC (linear)	70.77	64.58	65.88	64.41	58.72	48.78	70.90	64.53
SVC (RBF)	54.95	52.47	51.51	52.24	53.30	42.01	55.03	52.43
Decision Tree	55.05	52.69	52.31	51.62	53.35	43.72	53.97	51.27
MLP	83.80	81.83	82.18	82.87	65.05	66.55	84.61	84.49
LSTM	82.86	66.01	79.62	67.77	66.20	55.18	83.14	64.62
BiLSTM	81.75	67.78	81.94	65.64	68.89	54.53	83.05	65.40

Table 5: Results for the different combination of features comparing random data splitting to leaving out an individual speaker.

from a speaker, the task is more difficult. However, it is less obvious why this affects the MLP and the non-neural method (using MFCCs) only mildly but the LSTMs and the other features to a much higher degree. This will require a more in-depth analysis.

7 Conclusion and Future Work

In this paper, we investigated the efficacy of various audio input representations for accurately classifying discourse meaning in Spanish. We explored pitch contour representation using three audio features and compared the efficiency of utilizing means with different algorithms. We also evaluated if these features convey different information and their generalizability across speakers. Our findings suggest that using a combination of the three features with a recurrent neural network architecture provides the best results for our discourse-meaning classification task.

We also found that there is speaker specific information represented in the features, and that that combination of MLP and MFCCs is much more robust in a setting where we test on an unknown speaker than the other combinations. We will need to have a closer look to understand better why this is the case.

We are also planning on extending the corpus to include more speakers, and to balance it for dialects.

8 Limitations

It is important to explain the limitations of the current study. The corpus used for the experiment is limited in size and scope, which may have impacted the generalizability of the results. Further experiments with larger corpora that encompass a broader range of discourse meanings and linguistic features are necessary to validate and extend the findings of this research. Nevertheless, the present study provides valuable insights into the interaction between classifier and feature types, which will need to be considered in future experiments.

Acknowledgments

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Pablo Daniel Agüero and Antonio Bonafonte. 2004. Intonation modeling for TTS using a joint extraction and prediction approach. In *Fifth ISCA Workshop on Speech Synthesis*, Pittsburgh, PA.

Sankaranarayanan Ananthakrishnan and Shrikanth S Narayanan. 2005. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In *Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I/269–I/272, Philadelphia, PA.

Youakim Badr, Partha Mukherjee, and Sindhu Madhuri Thumati. 2021. Speech emotion recognition using MFCC and hybrid neural networks. In *Proceedings of the 13th International Joint Conference on Computational Intelligence (IJCCI)*, pages 366–373, Online.

Gérard Bailly and Bleicke Holm. 2005. SFC: A trainable prosodic model. *Speech Communication*, 46(3-4):348–364.

Gajanan K Birajdar and Mukesh D Patil. 2020. Speech/music classification using visual and spectral chromagram features. *Journal of Ambient Intelligence and Humanized Computing*, 11(1):329–347.

Eleanor Chodroff and Jennifer Cole. 2019. Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English. In *Proceedings of INTERSPEECH 2019*, pages 1966–1970, Graz, Austria.

Jennifer Cole, Jeremy Steffman, and Sam Tilsen. 2022. Shape matters: Machine classification and listeners' perceptual discrimination of American English intonational tunes. In *Proceedings of Speech Prosody*, pages 23–26, Lisbon, Portugal.

Namrata Dave. 2013. Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology*, 1(6):1–4.

Eva Estebas-Vilaplana and Pilar Prieto. 2010. *Transcription of Intonation of the Spanish Language*, chapter Castilian Spanish intonation. LINCOM.

Timothy L Face. 2004. The intonation of absolute interrogatives in Castilian Spanish. *Southwest Journal of Linguistics*, 23(2):65–80.

Timothy L Face. 2005. F0 peak height and the perception of sentence type in Castilian Spanish. *Revista internacional de lingüística iberoamericana*, 3(2 (6):49–65.

Timothy L Face. 2008. *The Intonation of Castilian Spanish Declaratives and Absolute Interrogatives*. Lincom Europa.

Timothy Lee Face. 2001. *Intonational Marking of Contrastive Focus in Madrid Spanish*. The Ohio State University.

Hiroya Fujisaki. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The Production of Speech*, pages 39–55. Springer.

Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. 2017. Speech intention classification with multimodal deep learning. In *Advances in Artificial Intelligence: 30th Canadian Conference on Artificial Intelligence (Canadian AI)*, pages 260–271, Edmonton, Canada.

Daniel Hirst and Robert Espesser. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15:75–85.

José Ignacio Hualde and Pilar Prieto. 2015. Intonational variation in Spanish: European and American varieties. In *Intonation in Romance*. Oxford University Press.

Dias Issa, M Fatih Demirci, and Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894.

Greg Kochanski and Chilin Shih. 2003. Prosody modeling with soft templates. *Speech Communication*, 39(3-4):311–352.

D Robert Ladd. 2008. *Intonational Phonology*. Cambridge University Press.

Gina-Anne Levow. 2005. Context in multi-lingual tone and pitch accent recognition. In *Ninth European Conference on Speech Communication and Technology*.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

Taniya Mishra, Vivek Rangarajan Sridhar, and Alistair Conkie. 2012. Word prominence detection using robust yet simple prosodic features. In *Thirteenth Annual Conference of the International Speech Communication Association*, Portland, OR.

Jinfu Ni, Hisashi Kawai, and Keikichi Hirose. 2006. Constrained tone transformation technique for separation and combination of Mandarin tone and intonation. *The Journal of the Acoustical Society of America*, 119(3):1764–1782.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Janet Pierrehumbert. 1981. Synthesizing intonation. *The Journal of the Acoustical Society of America*, 70(4):985–995.

Antonio Quilis. 1993. *Tratado de fonética y fonología españolas*. Gredos.

Andrew Rosenberg. 2009. *Automatic Detection and Classification of Prosodic Events*. Columbia University.

Andrew Rosenberg. 2010. Autobi - A tool for automatic toBi annotation. In *Eleventh Annual Conference of the International Speech Communication Association*.

Andrew Rosenberg. 2012. Modeling intensity contours and the interaction of pitch and intensity to improve automatic prosodic event detection and classification. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 376–381. IEEE.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Paul Taylor. 2000. Analysis and synthesis of intonation using the tilt model. *The Journal of the Acoustical Society of America*, 107(3):1697–1714.

Francisco Torreira and Simeon Floyd. 2012. Intonational meaning: The case of Spanish yes-no questions. In *the Fifth European Conference on Tone and Intonation (TIE5)*.

Jan PH Van Santen and Bernd Möbius. 2000. A quantitative model of F0 generation and alignment. In *Intonation*, pages 269–288. Springer.

Colin W Wightman and Mari Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481.

Lei Xie and Zhi-Qiang Liu. 2006. A comparative study of audio features for audio-to-visual conversion in MPEG-4 compliant facial animation. In *2006 International Conference on Machine Learning and Cybernetics*, pages 4359–4364. IEEE.

Albert Zeyer, Patrick Doetsch, Paul Voigtlaender, Ralf Schlüter, and Hermann Ney. 2017. A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2462–2466. IEEE.

Hengshun Zhou, Debin Meng, Yuanyuan Zhang, Xiaojiang Peng, Jun Du, Kai Wang, and Yu Qiao. 2019. Exploring emotion features and fusion strategies for audio-video emotion recognition. In *2019 International Conference on Multimodal Interaction*, pages 562–566.

A Parameters

Model	Parameter	Optimal setting
Random Forest	Estimators	500
	Criterion	entropy
	Warm Start	True
	Max Features	sqrt
	OOB Score	True
	Random State	69
KNN	Neighbors	5
	Weights	distance
	Algorithm	brute
	Leaf Size	30
	Jobs	30
SVC	C	10
	Gamma	auto
	Kernel	linear, rbf
	Random State	69
decision tree	Max depth	None
	Min sample leaf	2
	Min sample split	5
MLP	Activation	ReLU
	Alpha	0.0001
	Beta 1	0.9
	Beta 2	0.999
	Batch size	256
	Epsilon	1e-08
	Hidden Layer Sizes	(300,)
	Learning Rate	adaptive
	Solver	adam
(bi)LSTM	Layers	64
	Units	6
	Activation	softmax
	Learning Rate	0.01
	Optimization	adam
	Loss	Sparse Categorical Cross-entropy
	Batch Size	32
	Epochs	150

Table 6: Optimal parameter values for the models used in our experiments.