Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Combining Acoustic and Linguistic Features in Phrase-Oriented Prosody Prediction

Mónica Domínguez[1], *Mireia Farrús*[1], *Leo Wanner*[2,1]

[1]TALN Group, N-RAS Research Centre
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona, Spain
[2]Catalan Institute for Research and Advanced Studies (ICREA)
{monica.dominguez|mireia.farrus|leo.wanner}@upf.edu

Abstract

Intonation is traditionally considered to be the most important prosodic feature, whereupon an important research effort has been devoted to automatic segmentation and labeling of speech samples to grasp intonation cues. A number of studies also show that when duration or intensity are incorporated, automatic prosody labeling is further improved. However, the combination of word level acoustic features still attains poor results when machine learning techniques are applied on annotated corpora to derive intonation for speech synthesis applications. To address this problem, we present an experimental set-up for the development of a hierarchical prosodic structure model which combines linguistic features, including information structure, and three acoustic elements (intensity, pitch and duration). We show empirically that this combination leads to a considerably more accurate representation of prosody and, consequently, a more reliable automatic labeling of speech corpora for machine learning.

Index Terms: information structure, thematicity, prosodic label, prosodic phrase, prosodic word, ToBI, hierarchical prosodic structure, z-score, acoustic parameter.

1. Introduction

In speech synthesis, pitch[1] has been traditionally considered to be the most important prosodic feature, to be modelled in terms of variations of fundamental frequency (F0) at the word level. It has also been argued that generated speech quality is enhanced when prominence (salient peaks of F0) and phrasing (pause insertion) are complemented by lexical and syntactic [1–3] or information structure features [4] for phrase-based intonation generation. However, the combination of word level acoustic and linguistic features seem still not to lead to an optimal speech quality. Speech generated using these features continues to suffer from a certain monotony, especially in the case of expressive multisentential discourse [5].

To improve on intonation contour generation, Domínguez et al. [6] argue that, in accordance with Ladd and Selkirk's [7, 8] hypothesis of a hierarchical prosodic structure, intonation representation surpasses syllable and word boundaries to give way to intonation patterns at the intonation phrase (IP) level. The hierarchical prosodic structure foresees that prosody is segmented into different units, which are nested in a multi-layered hierarchical structure; as Selkirk's *Strict Layer Hypothesis* [8] describes.

Domínguez et al. [6] show that intonation can indeed be better correlated with information structure (IS) in the context of the Information Structure – Prosody interface [4, 9] at the IP level than at the word level, and thus lead to more accurate intonation contours. But since their exploration was limited to a mere correlation between information structure and ToBI labels [10], it remained unclear how other acoustic and linguistic features may influence the accuracy of the prediction of prosody.

In what follows, we set out to assess whether the accuracy of the prediction of prosodic labels at the prosodic phrase (PPh) level is further improved when three acoustic, namely intensity (dB), pitch (Hz) and duration (ms), and linguistic elements (word position, syntax, and information structure) are taken into account. In this respect, we use the terms *prosodic label* and *prosodic phrase* implying the integration of several acoustic traits (so far, intensity, pitch and duration) as opposed to *intonation label* and *intonation phrase*, which in speech technologies mainly involve the manipulation of pitch, understood as the variation of F0 values along the utterance.

The remainder of the paper is structured as follows. The next section describes the dataset that we use for our experiments. In Section 3, we show that acoustic parameters also correlate in terms of prosodic labels at the phrase and sentence level and therefore cannot be neglected. Then, we combine the acoustic and linguistic levels (including information structure) and demonstrate in classification experiments that, together, they lead to a more accurate prediction of prosodic labels. Moreover, our experiments suggest that some general characteristics are maintained across speakers, even if each speaker varies in the relative amount of each parameter used when expressing the same communicative content. Section 4 then provides empirical data supporting the hypothesis that prosodic labels at the PPh layer are characterized in terms of a distinct combination of intensity, pitch and duration in standard deviation values, Section 5 offers some conclusions and outlines aspects of our future work in this area.

2. The Dataset

We use a fragment of 109 sentences from the Wall Street Journal (WSJ) Penn Treebank corpus [11] as our working corpus. The selection of this set of sentences was made considering a varied syntactic composition which allows representativeness

[1]The term *pitch* in this paper refers to one of the three acoustic elements, being the other two intensity and duration. Pitch comprises in our experiments the following set of acoustic parameters: minimum F0, maximum F0 and F0 range.

 doi: 10.21437/SpeechProsody.2016-163

of information structure parameters in terms of: (i) the amount of hierarchical levels of thematicity (up to three in the data set); (ii) the presence/absence of each communicative span; (iii) their position within the sentence and with respect to each other; (iv) and the spans' continuity or lack of it.

The corpus has been processed using Bohnet's [12] joint tagger and dependency parser to obtain lexical and syntactic features and annotated manually with information structure (more precisely, with Thematicity[2] features from Mel'čuk's [9] *communicative structure*), following the guidelines established by Bohnet et al. [13]. Further details on how information structure is understood can be found in the authors' work [6] and [14] on its correlation to prosody.

Twelve native speakers of American English were recruited for a recording session in a professional studio and they were asked to read each sentence independently. Segmentation of the audio files and extraction of acoustic values has been automatically carried out using Praat [15]. Extracted acoustic parameters include: (i) intensity (minimum, maximum and average) in dB, (ii) pitch (minimum, maximum and range) in Hz, and (iii) duration (per word interval) in ms.

The subsequent manual annotation of prosodic labels for the creation of the training set relied on the combination of at least two salient acoustic parameters at the sentence level. During the annotation, words have been labelled as lexically stressed (S) and unstressed words (U) if they are not salient at sentence level. On the other hand, words which are prosodically prominent have been labelled as pitch accents (PA) or boundary tones (BT), each of which may take one of the possible ToBI labels shown in Table 1. We are using the reduced set of prosodic labels specified in Table 1 including six ToBI labels and two prosodic marks (S and U) for our classification problem.

Table 1: Prosodic Annotation Schema

Prosodic Mark	Prosodic Label
PA	H*
	L*
	L*+H
BT	HL%
	LL%
	LH%
S	S
U	U

In the next section, we present the results from applying this annotation schema in our corpus to a collection of classification experiments that explore the combination of the acoustic and linguistic levels using all speakers and individual speakers.

3. Prosody Prediction Experiments

For our experiments on prosody prediction, we used Weka's J48 tree classifier [16], with a 10 fold cross-validation. In order to account for the linear nature of our classification problem, a time series filter (to use Weka's terminology) has been applied to all features. That is, the preceding prosodic label (n-1) and all its features are used to predict the pending prosodic label n. Accuracy, kappa, and root mean square error (RMSE) are used to assess each feature's performance.

Three experiments have been carried out. The first two served to assess the potential of acoustic parameters and linguistic features to predict prosodic labels using all speakers' voice samples. For this purpose, we use each acoustic element (pitch, intensity and duration) separately, and then, in combination. Each element includes the corresponding set of acoustic parameters specified in the previous section. In the second experiment, communicative (i.e., information structure) and morpho-syntactic features are combined with acoustic parameters. In the third experiment, we explore to what extent acoustic parameters are speaker dependent in order to validate the practice of the state-of-art technologies to use samples obtained from different speakers for training prosody models.

3.1. Testing acoustic elements

Table 2 shows that when we use pitch, intensity and duration on their own to predict prosodic labels, the best results are achieved by duration (53.3%) and the lowest by pitch (41.8%). The highest performance, however, is achieved by the combination of all three elements: intensity, duration and pitch (56.3%). In other words, prosody is realized by a combination of at least these three prosodic elements based upon observation on our speech corpus. These results suggests that, as already pointed out by Audibert et al. [5], in order to obtain a more natural synthesized voice, more acoustic elements (not only pitch) should be integrated.

Table 2: Classification Results. Acoustic Parameters

Acoustic Parameters	Accuracy	Kappa	RMSE
Pitch	41.8%	0.26	0.32
Intensity	42.1%	0.26	0.32
Duration	53.3%	0.40	0.28
Pitch + Intensity	45.5%	0.31	0.32
Duration + Intensity	53.8%	0.42	0.30
Duration + Pitch	55.3%	0.43	0.29
Intensity + Duration + Pitch	**56.3%**	**0.48**	**0.29**

Moreover, this suggests that an automatic prosody labeler will perform better when three acoustic elements are integrated (56.3%), instead of using only pitch and duration (55.3%), as AuToBI does [17].

3.2. Adding linguistic features

Following the argumentation in [1–3] that adding linguistic features to the acoustic parameters helps improve prosody prediction, we added linguistic features to the acoustic parameters of the first experiment. Table 3 presents results from combining the acoustic level with each linguistic element: word position, syntax and thematicity; and the linguistic level with each acoustic element (intensity, pitch and duration).

We see that, compared to the first experiment (where a maximum of 56.3% was achieved), prediction accuracy improved considerably when the acoustic level is enriched by each linguistic element. The syntactic element is particularly useful (leading to an accuracy of 71.7%). The picture improves even further when the linguistic level is taken as basic features and individual acoustic elements (or a combination thereof) are used

[2] Mel'čuk's [9] Thematicity features are *theme*, *rheme* and *specifier*. In contrast to the traditional bipartite Thematicity [4], Mel'čuk's thematicity is hierarchical, which accounts for embeddedness of communicative spans, and is consequently instrumental for complex sentence generation and correlation of prosodic contours at different prosodic layers.

Table 3: Classification Results. Acoustic and Linguistic Features

Features		Accuracy	Kappa	RMSE
Acoustic +	Position	58.7%	0.48	0.28
	Thematicity	60.1%	0.49	0.27
	Syntax	71.7%	0.64	0.23
Linguistic +	Intensity	76.6%	0.70	0.21
	Duration	76.6%	0.70	0.20
	Pitch	**77.9%**	**0.72**	**0.20**
All features		72.5%	0.65	0.22

to complement them. As can be observed, a combination with pitch reaches 77.9%.

In other words, the combination of acoustic and linguistic features improves the prediction accuracy up to nearly 23%. However, the best result is achieved by the combination of the linguistic level with each acoustic element separately, rather than by the combination of all linguistic and acoustic features (when 72.5% is reached). We hypothesize that this is because in this experiment we draw upon the dataset of all speakers for training. Therefore, in the next subsection, we explore whether the picture changes when we work with speaker-specific datasets.

3.3. Speaker dependency of acoustic parameters

Five individual speakers have been compared in order to test results of the previous experiments on all speakers samples. The selected speakers belong to different dialectal regions in the USA (New York, Illinois, Texas, Boston and Arizona) represented in blue in Figure 1.

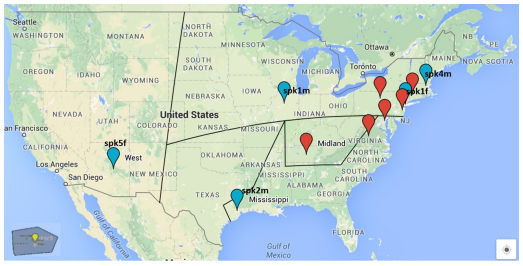

Figure 1: Speakers' state of birth

Figure 2 contrasts for each speaker-specific sample the prediction accuracy of the combination of the linguistic level with each acoustic element individually against the combination of the linguistic level with all three acoustic elements together. It is shown that indeed, for each individual speaker, the combination of the linguistic level with three acoustic elements leads to a higher performance than a combination with only one acoustic element. The overall prediction quality decreases from 72.5% when the samples of all speakers are used to 63% (as the maximum accuracy obtained by individual speakers) when the voice sample of only one speaker is used. This is because the amount of training data decreases in each individual case to a point where it is insufficient for optimal training. But this shortcoming can be easily overcome using bigger datasets of speaker-specific data.

On the other hand, it is remarkable that when each of the acoustic elements is combined with the whole linguistic level,

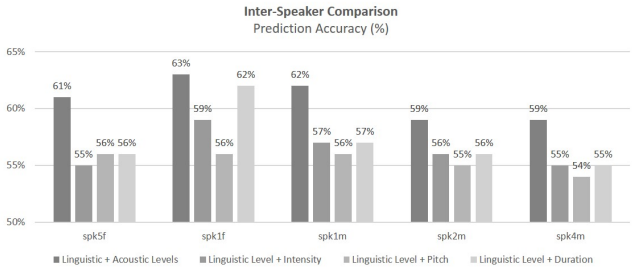

Figure 2: Comparison of inter-speaker accuracy

similar prediction accuracy (around 56%) is obtained for each speaker. However, with some speakers (i.e. spk1f) higher accuracy is achieved for a specific acoustic element (in this case, duration). For all speakers, the combination of both linguistic and acoustic levels leads, once again, to better accuracy results (average of 61%) than when using only one of the acoustic elements.

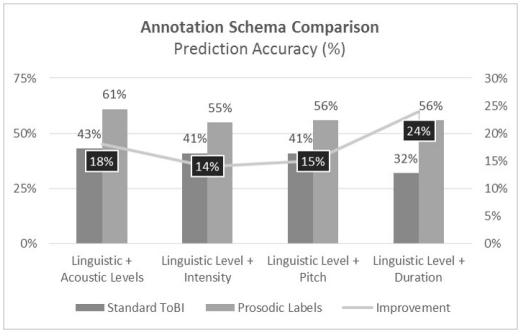

Figure 3: Annotation Schema Comparison

One speaker (spk5f) was considered for testing the improvement of our annotation schema versus a detailed description of the intonation contour using a detailed annotation with ToBI labels carried out by an expert annotator. Figure 3 shows a prediction accuracy improvement for all combinations of acoustic elements with the linguistic level when a reduced annotation schema over prosodic labels is used. Further research on inter-annotator agreement using each system was out of the scope of this paper. The improvement (labeled in a black box in Figure 3) goes from 14% to 24% of prediction accuracy surpassing the 50% accuracy level when our prosodic labeling annotation schema is used for one speaker.

Accuracy is expected to be much higher when a larger dataset for training is used. But, still, we must keep in mind that each speaker may chose a different combination of acoustic parameters to mark prosodic prominence and phrasing in a different way. Consequently, using different speakers for training may result in classification shortcomings if data is not treated independently for each speaker.

4. Acoustic Elements at the Prosodic Phrase Layer

We have shown that combining acoustic elements with linguistic features considerably enhances the prosody prediction potential of a model using prosodic labels, compared to de-

tailed ToBI labeling, especially if we consider speaker-specific voice samples. Let us now explore the characterization of each prosodic label for each speaker in terms of the combination of acoustic parameters that can be used both to annotate and generate prosody automatically.

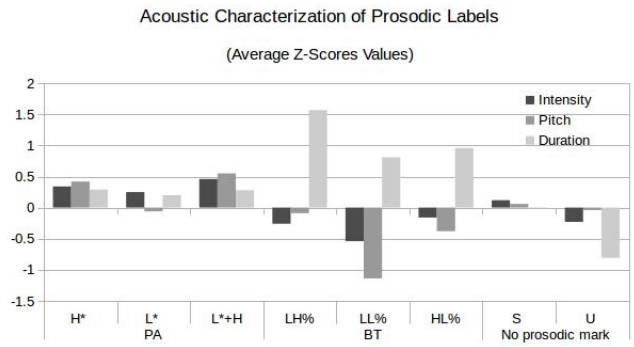

Figure 4: Characterization of prosodic labels combining acoustic elements

As already mentioned in the Introduction, Ladd [7] and Selkirk [8] establish a universal hierarchy of prosodic constituents. Our approach considers the *prosodic phrase* (PPh) and *prosodic word* (PWd) as two intermediate layers between the utterance and the syllable. These prosodic units are inclusive for three acoustic elements, opposed to using only pitch as the most relevant prosodic feature to be modeled at a word level, as traditionally done in Speech Technologies. In what follows, we aim to describe prosody as a vectorial representation of acoustic parameters at the PPh layer. For this purpose, we normalize absolute values of the features using z-scores for each speaker and sentence. Moreover, the average z-score for each acoustic element is calculated to observe prosodic labels characteristics in terms of standard deviation.

Figure 4 shows that both prosodic labels and prosodic marks (PA, BT, S and U) are represented as a distinct combination of intensity, pitch and duration elements. Thus, PAs (H*, L*, L*+H) are characterised by positive or null deviation in all three acoustic elements, while words labelled as 'S' have little or no deviation at all. BTs, on the other hand, are characterized by a high positive deviation in duration and negative deviation in intensity and pitch. Finally, words labelled as 'U' have outstandingly low negative deviation in duration and negative deviation in intensity and duration. Summing up, Figure 4 proves that predicting and deriving prosodic labels from acoustic elements systematically is feasible since prosodic labels show a characteristic combination of acoustic parameters. Moreover, the advantage of working at the PPh level is that it can serve as a scaffolding upon which expressiveness can be constructed in a more traceable way when moving across prosodic layers.

5. Conclusions

In this paper, we presented a prosody model that combines acoustic and linguistic features and takes into account each speaker's voice sample at the PPh layer. Such an approach reflects better communicative expressiveness of natural speech, and is, thus, more appropriate in expressive speech generation for text-to-speech (TTS) and content-to-speech (CTS) applications. Furthermore, the approach introduced in this paper involves a linguistically and acoustically traceable and theoretically motivated prosodic annotation schema which facilitates the compilation of training datasets.

A further asset of this work is the fact that using machine learning approaches allows feature engineering experiments that can be used as empirical evidence to advance greatly in the research area of Speech Prosody. Furthermore, objective annotation schemas based upon numeric values of acoustic parameters, like the one we presented in this paper, are essential to analyze, automatically label and explore large amounts of data.

Finally, if we follow theoretical studies that consider different prosodic layers nested in a hierarchical structure, the next prosodic layer (i.e., the prosodic word) needs to be predicted upon the results obtained at the prosodic phrase. Further research is being carried out in this direction using a corpus of spontaneous speech to establish a sound correlation between PPh and PWd layers following the methodology presented in this paper. The goal is to implement this model and evaluate it in an automatic prosodic labeling system to train a prosodic module for generating expressive prosody in CTS applications.

6. Acknowledgements

This work is part of a project that has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under the Grant Agreement number H2020-RIA-645012. The second author is partially funded by a grant from the Spanish Ministry of Economy and Competitivity in the framework of the *Juan de la Cierva* fellowship program.

7. References

[1] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[2] A. Schweitzer and B. Möbius, "Experiments on automatic prosodic labeling," in *Proceedings of Interspeech 2009 (Brighton)*, 2009, pp. 2515–2518.

[3] A. Wagner, "Automatic Labelling of Prosody," *Speech and Language Technology*, vol. 14/15, no. special edition, 2011.

[4] M. Steedman, "Information structure and the syntax-phonology interface," *Linguistic inquiry*, 2000.

[5] N. Audibert, V. Aubergé, and A. Rilliard, "LNCS 3784 - The Relative Weights of the Different Prosodic Dimensions in Expressive Speech: A Resynthesis Study," *ACII 2005, LNCS 3784*, pp. 527–534, 2005.

[6] M. Domínguez, M. Farrús, A. Burga, and L. Wanner, "The Information StructureProsody Language Interface Revisited," in *Proceedings of the 7th International Conference on Speech Prosody (SP2014)*, Dublin, Ireland, 2014, pp. 539–543.

[7] R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 2008.

[8] E. O. Selkirk, *Phonology and Syntax: The relation between sound and structure*, 1984.

[9] I. Mel'čuk, *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*, 2001.

[10] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "TOBI: A Standard for Labeling English Prosody," in *2nd Internatioanl Conference on Spoken Language Processing (ICSLP 92)*, no. October, 1992, pp. 867–870.

[11] E. Charniak and E. al., "BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43." Philadelphia, 2000.

[12] B. Bohnet, A. Björkelund, L. Hafdell, and P. Nugues, "A high-performance syntactic and semantic dependency parser," *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations (COLING 2010)*, no. August, pp. 33–36, 2010.

[13] B. Bohnet, A. Burga, and L. Wanner, "Towards the Annotation of Penn TreeBank with Information Structure," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, 2013, pp. 1250–1256.

[14] M. Domínguez, M. Farrús, A. Burga, and L. Wanner, "Using Hierarchical Information Structure for Prosody Prediction in Content-to-Speech Applications," in *International Conference on Speech Prosody (SP2016), Boston, USA*, 2016.

[15] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2015. [Online]. Available: retrieved 21 february 2015 from http://www.praat.org/

[16] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann, 2005.

[17] A. Rosenberg, "AutoBI-a tool for automatic toBI annotation." *INTERSPEECH*, 2010.