Prosodic Features' Criterion for Hebrew

Ben Fishman[1], Itshak Lapidot[2], and Irit Opher[2](✉)

[1] Tel Aviv University, Tel Aviv, Israel
benf22@gmail.com
[2] Afeka Academic College of Engineering, Tel Aviv, Israel
{itshakl,irito}@afeka.ac.il

Abstract. Prosody provides important information about intention and meaning, and carries clues regarding dialogue turns, phrase emphasis and even the physiological or emotional condition of the speaker. Prosody has been researched extensively by linguists and speech scientists; However, little attention has been given to formulating and ranking the acoustic features that represent prosodic information. This paper aims at defining a simple methodology that allows us to test whether a feature conveys prosodic information. This way, we can compare different features and rate them as prosodic or content related (In this paper the word "content" refers to the verbal information of the utterance.). We explore many features using a Hebrew dataset especially designed for validating prosodic features, and as the first step of our research we chose two prosody classes: neutral and question. We apply our methodology successfully and find that prosodic features indeed are invariant to the content of the utterance, while correlating with prosodic manifestations. We validate our methodology by showing that our ranking of prosodic features yields similar results to classification based feature selection.

Keywords: Prosody · Prosodic features · Hebrew database

1 Introduction

Prosody can be defined as the study that relates to non-contextual aspects of speech. Prosody provides valuable information that can be perceived by the listener and plays an important role in everyday life – it helps maintaining dialogue structure [6], decyphering higher level utterances (e.g. sarcasm) and assessing the speaker's emotional state, attitude or intentions [13]. It also contributes to the medical field, especially in Neurology [4,12]. Prosody is also essential for many speech based systems, such as Text to Speech (TTS) [2], Speech Morphing [14] or Speech based Analysis [11].

In the past years there has been extensive work towards standardization of prosody transcription, for example ToBI [15] that is used for annotating intonation or the IPrA Prosodic Alphabet [7]. Still, little attention has been given so far to generalizing and formulating the acoustic features that represent the perceived

P. Sojka et al. (Eds.): TSD 2018, LNAI 11107, pp. 482–491, 2018.
https://doi.org/10.1007/978-3-030-00794-2_52

intonation and other prosodic building blocks, especially in under resourced languages such as some of the Semitic languages.

In this work we take a few steps towards such a formulation and define a simple methodology for determining whether an acoustic or spectral feature represents prosodic information and to what degree.

2 Features

Prosodic features are widely used for various tasks – emotion detection [1], language identification [16], TTS [2], etc'. There has been extensive work on extracting various acoustic and spectral features for prosodic research, e.g. [17]. The openSMILE project [5] lists a few hundred features for emotion recognition. Other works, such as [3] limit themselves to features that can be derived from F0, duration and energy only, that are the most commonly used prosodic features. So, there are many features one can use, but is there a way to decide whether a feature indeed carries prosodic information? Our suggested methodology tries to address this issue and is presented in Sect. 3. To demonstrate and test this methodology, we use a 48 feature set, most of which are considered standard in prosodic research, e.g. F0 and its derivatives, while some are hand-crafted features such as amplitude-tilt ($A = \frac{|A_r|-|A_f|}{|A_r|+|A_f|}$, when A_r/A_f is the amount of F0 rise/fall respectively) or duration-tilt ($D = \frac{|D_r|-|D_f|}{|D_r|+|D_f|}$, where D_r/D_f is the F0 rise/fall duration respectively) [11]. All features are listed in Table 1.

Naturally, these features can be scalars, e.g. max value of F0 or vectors e.g. average energy per syllable, or MFCC entries per frame. To obtain syllable boundaries, we used word level forced-alignment using Hebrew acoustic models trained with the Kaldi engine.

Table 1. Feature set list

Directly calculated	Derived features	Segments types	Num
F0, dF0, energy	Max, min, mean, var	Per-syllable, accumulated	24
F0, dF0	Max-range		4
F0	Peak-position, ampTilt, durTilt		6
MFCC		Per-frame	13
Duration		Per-syllable	1

Per-syllable: evaluated over a single syllable. Accumulated: evaluated over a segment starting at the beginning of the syllable and ending at the end of the utterance.

3 Methodology for Evaluating a Prosodic Feature

We wish to define a criterion for measuring how well does a feature represent prosodic information conveyed in speech utterances. The proposed methodology is simple and requires the following: (1) prosodic features should be correlated with some prosodic manifestations and (2) should not be correlated with significant changes in other "dimensions", when the same prosody is used. Note that requirement (2) relates in our paper only to the content of the utterance, as we use an over simplification where only two attributes characterize a feature – prosodic or content related. Naturally this does not hold for some languages such as tonal languages, where different tonal patterns convey content [8] hence these languages will have to be considered separately.

3.1 Formulation

Let us formulate the above requirements: Suppose we have a set of utterances U_{pc}^{k}, where $p = 1, 2, \ldots, N_P$ is an index representing the different prosodies in our dataset, $c = 1, 2, \ldots, N_C$ represents the different content types, i.e. different phrases, and $k = 1, 2, \ldots, K_{pc}$ runs through all utterances of type pc, i.e. all the utterances in our dataset with prosody p and phrase type c. Feature F is denoted prosodic if the following requirements hold:

Requirement 1: The dissimilarity between the extracted features is sufficiently small, for most utterance pairs with the same prosody p: $d_F\left(F_{pc}^{k}, F_{pr}^{l}\right) < T_1$ for more than x_1% of the pairs, where $d_F(\cdot, \cdot)$ is the dissimilarity between two features, and T_1 is some threshold we use to define "sufficiently small". x_1 and T_1 can be tuned, and d_F is defined for all features taking into account feature type – scalar or vector.

Requirement 2: The dissimilarity between extracted features for utterances with different prosodies is higher than T_2 for most such utterance pairs, $(q \neq p)$: $d_F\left(F_{pc}^{k}, F_{qc}^{l}\right) > T_2$ for more than x_2% of the pairs. This requirement should hold for both cases of same or different content type, but it is naturally stronger when the content is unchanged.

T_1, T_2, x_1, x_2 can be tuned for each feature and for each dataset. Instead of tuning these parameters, we propose to combine the two requirements in the following way – we require that the PMFs of the dissimilarities of the two sets (same prosody and different prosody) will be well separated reflecting the different behavior of the feature's values for the two sets; This means that we require low values of dissimilarities for the same prosody set, and high values of dissimilarity for the different prosody set.

3.2 Proposed Methodology

Our proposed methodology is depicted in Fig. 1 using a block diagram to determine the nature of the feature: (i) Calculate dissimilarities between feature values over all possible utterance pairs (ii) Group all pairs into two sets – "same prosody" set (S_{same}^{P}) and "different prosody" set (S_{diff}^{P}) (iii) For each set, evaluate the Probability Mass Function (PMF) using the normalized histograms of the

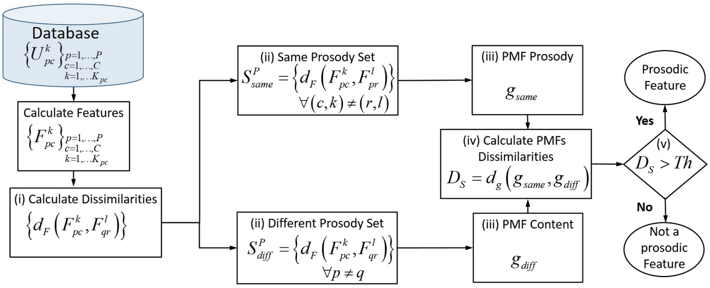
Database
$\left\{U^k_{pc}\right\}_{\substack{p=1,\ldots,P\\c=1,\ldots,C\\k=1,\ldots K_{pc}}}$
Calculate Features
$\left\{F^k_{pc}\right\}_{\substack{p=1,\ldots,P\\c=1,\ldots,C\\k=1,\ldots K_{pc}}}$
(i) Calculate Dissimilarities
$\left\{d_F\left(F^k_{pc},F^l_{qr}\right)\right\}$
(ii) Same Prosody Set
$S^P_{same}=\left\{d_F\left(F^k_{pc},F^l_{pr}\right)\right\}$ $\forall(c,k)\neq(r,l)$
(ii) Different Prosody Set
$S^P_{diff}=\left\{d_F\left(F^k_{pc},F^l_{qr}\right)\right\}$ $\forall p\neq q$
(iii) PMF Prosody
g_{same}
(iii) PMF Content
g_{diff}
(iv) Calculate PMFs Dissimilarities
$D_S=d_g\left(g_{same},g_{diff}\right)$
(v)
$D_S>Th$
Yes
No
Prosodic Feature
Not a prosodic Feature

Fig. 1. Flow chart describing the proposed methodology for evaluating the prosodic nature of a feature

dissimilarities (g_{same} and g_{diff}) (iv) Calculate dissimilarity score, denoted D_s, between the two PMFs (v) Use a threshold Th to decide whether the feature can be considered prosodic, i.e. conveys prosodic information. The threshold Th reflects the requirement stated above regarding large enough separability between dissimilarities PMFs of the two sets – same prosody set and different prosody set.

In this work we use the Euclidean distance as the dissimilarity function $d_F(\cdot)$ for step (i) and the symmetrized KL-divergence as the function $d_g(\cdot)$ for step (iv) when calculating D_s. If the feature is indeed prosodic, we expect to see a high degree of separability between the two dissimilarities PMFs evaluated for the two sets – "same prosody" and "different prosody". If, on the other hand, we find that there is a high similarity between the two dissimilarities PMFs (g_{same} and g_{diff}), we conclude that this feature does not carry any prosodic information, at least for the prosodies that were used.

4 Datasets

We have used two different datasets:

4.1 Hebrew Dataset

Our main dataset[1] is in the Hebrew language, and was designed specifically for prosody research. As this was a preliminary stage, we used only two prosody types: question and neutral. 36 speakers were recorded (males: 47%, females: 53%) of various ages (20–30: 22%, 30–40: 33%, 40–50: 8%, 50–60: 20%, 60–70: 17%). Each speaker uttered the same three short phrases, that consisted of four syllables each. All phrases were syntactically correct, and contained mostly voiced phonemes. Each phrase was recorded in two different prosodies (neutral: 46%, question: 54%). The data was recorded using personal cellular phones, in

[1] Hebrew dataset is freely available for research purposes only, by contacting the authors.

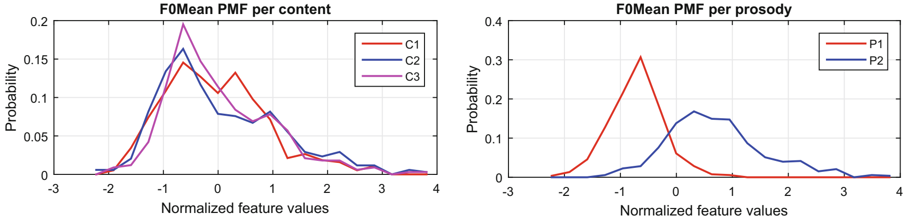

Fig. 2. Normalized F0Mean PMFs. Right – good separation between prosody classes. Left – no separation between content classes

a quiet room environment. In total there are 252 short phrases. To validate the data, two experienced listeners tagged all utterances in a random blind test. The manual tagging was 97% correct, therefore we consider the prosody labeling to be accurate.

4.2 Validation Dataset

In order to validate our results obtained with the Hebrew dataset, we used a small subset of LDC2002S28 – "Emotional Prosody Speech and Transcripts" corpus [9]. This English dataset contains recordings of professional actors reading a series of semantically neutral utterances using different emotional categories. We used a single speaker and five emotional categories only (anxiety, boredom, sadness, panic, and elation). This dataset is relevant for our research as prosody is often used to expresses emotional states.

5 Data Analysis

Based on our highly simplified approach, where acoustic features can be associated with either prosody or content, the first clue regarding the nature of the feature is evident when we look at the features' PMFs for different dataset partitionings – according to prosodies and according to content. Figure 2 shows an example for F0Mean, evaluated for each syllable in the Hebrew dataset. This feature seems to be prosodic – the feature values PMFs are significantly different for the prosody tagging, while content tagging yields similar PMFs. Indeed, F0Mean is considered to convey prosodic information. In Fig. 3 we can see the opposite behavior for one of the MFCC features (which are indeed more suitable for representing content).

Next we look at averages of features over time to explore the separation ability of a feature. In Fig. 4 we can see the behavior of two features over time. First we comapre between different prosodies and different contents for the Herbrew set. It is evident that DurTilt is prosodic while it does not separate well between our different phrases. We also look at F0Max for the validation dataset that includes five different prosody classes, and see that when using this single feature, it is possible to distinguish between some of these classes but not between all of them. This behavior indicates that naturally, F0Max carries prosodic information.

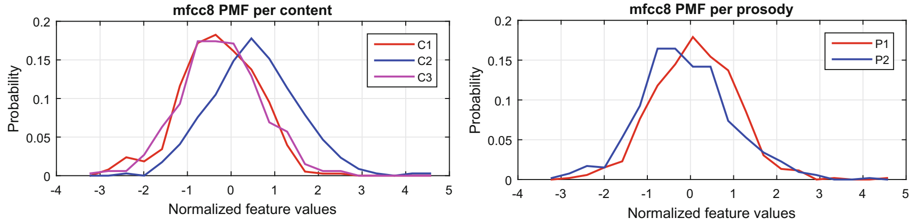

Fig. 3. Normalized MFCC8 PMFs. Right – no separation between prosody classes. Left – some separation between content classes

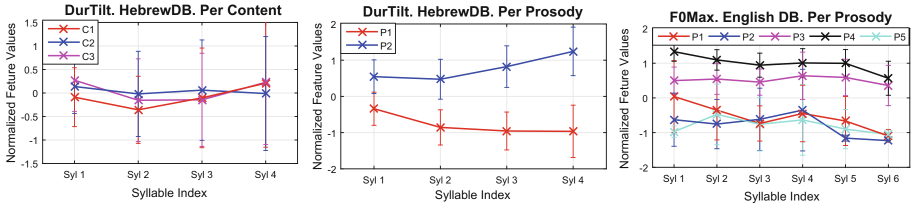

Fig. 4. Averaged feature values over syllables. Left & center: Hebrew dataset tagged per-content and per-prosody. Right: English dataset tagged per-prosody

5.1 Dissimilarities

Following steps (ii) and (iii) in our methodology, Fig. 5 shows the dissimilarity PMFs for the different sets – "same prosody" and "different prosody" for the duration-tilt feature for the Hebrew dataset. We can see good separation between the sets of same and different prosody, while there is no separation between the sets of same and different content. According to our criterion, this feature is definitely a "prosodic feature".

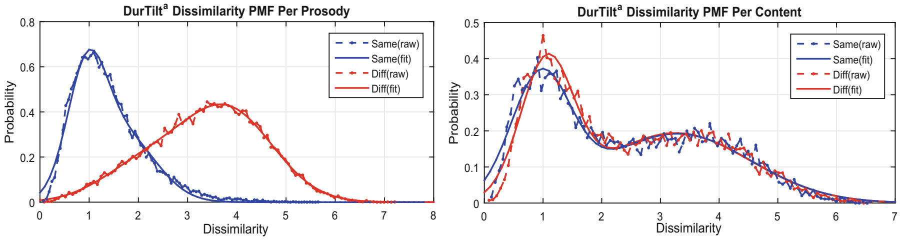

Fig. 5. Duration-tilt dissimilarity PMFs for different conditions: same and different prosodies, and same and different content

In Fig. 6 we can see another example of the dissimilarity PMFs between same and different prosody sets, for F0Max over the validation dataset. This feature separates well between prosody P4 (Panic) and P1 (Anxiety), while it does not separate between P4 and P3 (Elation). This feature also separates well between P4 and P2 (Boredom) and P5 (Sadness) (not shown).

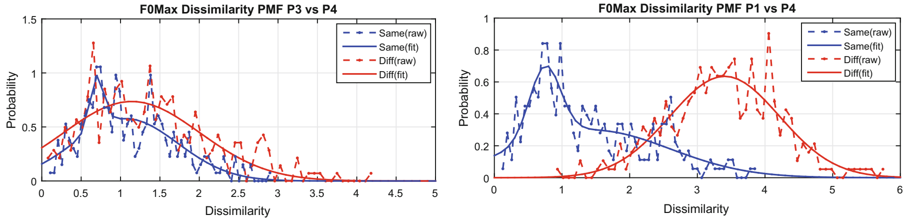

Fig. 6. PMFs dissimilarity for same and different prosody classes

5.2 Classification

One of our goals in grading and estimating feature quality is choosing the best features to be used in a classification task. Hence, to validate our proposed methodology we compare it with classifier based feature analysis. Since we currently defined only a single feature's ranking, we compare our results to ranking based on 1D classification results, i.e. classification using a single feature. We use a simple classifier, so for each feature separately we trained a logistic regression classifier using 66% of our data for training, while making sure train and test sets did not contain the same speakers. For each feature, we used a threshold that yields the best F_1 measure[2] over the train set. Applying this score to the test set, we obtained classification accuracy for the test set. These accuracy scores were used to rank the features for the classification task.

Next, we chose the 14 highest ranked feature according to our methodology, i.e. with highest D_s score, and compared them with the 14 best classification features i.e. the features that yielded the highest F_1 scores. We found that 13 features appeared in both lists (see Table 2), some of them with similar ranks.

Table 2. Comparison between ranking produced by the proposed methodology (D_s) and by classification (F_1). X^a denotes accumulated features as explained in Table 1

Feature	AmpTilt^a	DurTilt^a	AmpTilt	DurTilt	F0Mean^a	dF0Mean^a	F0Max^a
D_s	11.24	10.28	6.13	5.21	4.33	3.69	2.23
F_1	0.89	0.88	0.78	0.78	0.81	0.89	0.82
Feature	F0Mean	F0Max	dF0Max^a	dF0Mean	dF0Max	F0Range^a	F0Var^a
D_s	1.98	1.59	1.3	0.85	0.73	0.71	0.52
F_1	0.72	–	0.83	0.73	0.74	0.77	0.82

[2] F_1 measure is the harmonic mean of recall and precision: $F_1 = \frac{2}{1/recall+1/precision}$.

5.3 Dimensionality Reduction

When using more than one feature, we need to visualize this high-dimensional data and check separability between different classes. This can be done by applying dimension reduction schemes. We chose the t-SNE algorithm [10] and applied it to the best 15 prosodic features obtained using our D_s score (as explained in Sect. 3). Figure 7 shows very good separation between the two prosodic classes, while there is no separation at all between the content classes. This shows that the 15 best features do not represent the phrases content, in addition to conveying prosodic information. When repeating this process for the best content related features, we do not get any separation between different prosodies.

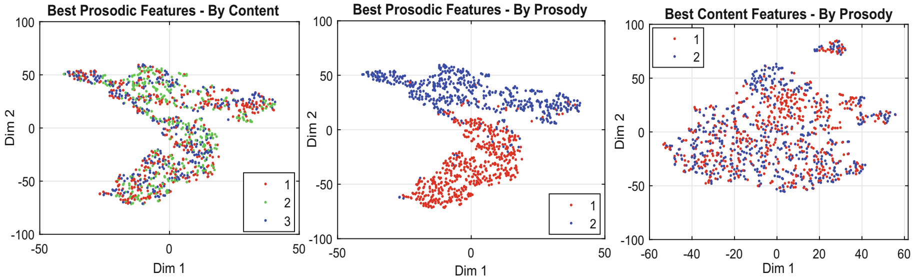

Fig. 7. 2D representation. Left & center – best prosodic features colored by content and by prosody. Right – best content features colored by prosody

6 Conclusions and Future Work

In this paper we introduced a methodology for validating the prosodic relevance of an acoustic or spectral feature. We refer to a feature as prosodic if its values differ significantly for utterances spoken with different prosodies, and show little or no change for utterances spoken with the same prosody. This methodology can be further extended to provide an estimation as to the "prosodic ranking" of a feature, taking into account its separation ability for different prosodic classes, and its insensitivity to changes in content and other non-prosodic information. We believe that creating a formal standard ranking mechanism for prosodic features can assist in finding representations for known perceptual notations such as IrPA or ToBI, as well as in revealing new prosodic features. This methodology can also be used for analyzing prosodic features and manifestations in different languages. For our Hebrew dataset, we have successfully shown that features that were ranked high based on our methodology, are indeed relevant for conveying prosodic information. This was done by ranking the single features according to their classification accuracy scores and comapring this ranking to the one induced by our proposed prosodic score.

Future work should address a few issues that were not covered: (1) Extending the methodology to deal with: a. tonal languages b. more than two prosody classes c. additional non-prosodic dimensions other than content (2) Providing a full mathematical formalization for D_s (3) Validating the proposed methodology using larger datasets in additional languages, as well as for other prosody classes, using known feature sets such as openSMILE [5] (4) Dealing with multi-feature classification results.

Acknowledgments. The authors thank Ella Erlich, Ruth Aloni-Lavi and Noga Hellman for their help with the Hebrew dataset.

References

1. Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Seventh International Conference on Spoken Language Processing (2002)
2. Chen, S.H., Hwang, S.H., Wang, Y.R.: An RNN-based prosodic information synthesizer for mandarin text-to-speech. IEEE Trans. Speech Audio Process. **6**(3), 226–239 (1998)
3. Rose, R.C.: Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. In: Sixth European Conference on Speech Communication and Technology (1999)
4. Diehl, J.J., Paul, R.: The assessment and treatment of prosodic disorders and neurological theories of prosody. Int. J. Speech-Lang. Pathol. **11**(4), 287–292 (2009)
5. Eyben, F., Wöllmer, M., Schuller, B.: OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462. ACM (2010)
6. Hastie, W.H., Poesio, M., Isard, S.: Automatically predicting dialogue structure using prosodic features. Speech Commun. **36**, 63–79 (2002)
7. Hualde, J., Prieto, P.: Towards an international prosodic alphabet (IPrA). Lab. Phonol. **7** (2016)
8. Li, S., Wang, Y., Sun, L., Lee, L.: Improved tonal language speech recognition by integrating spectro-temporal evidence and pitch information with properly chosen tonal acoustic units. In: INTERSPEECH (2011)
9. Liberman, M.: Emotional Prosody Speech and Transcripts LDC2002S28 (2002). https://catalog.ldc.upenn.edu/LDC2002S28
10. Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
11. Mary, L., Yegnanarayana, B.: Extraction and representation of prosodic features for language and speaker recognition. Speech Commun. **50**(10), 782–796 (2008)
12. McCann, J., Peppé, S.: Prosody in autism spectrum disorders: a critical review. Int. J. Lang. & Commun. Disord. **38**(4), 325–350 (2003)
13. Pierre-Yves, O.: The production and recognition of emotions in speech: features and algorithms. Int. J. Hum.-Comput. Stud. **59**(1–2), 157–183 (2003)
14. Qavi, A., Khan, S.A., Basir, K.: Voice morphing based on spectral features and prosodic modification. In: Multi-Topic Conference (INMIC), pp. 401–405. IEEE (2014)
15. Silverman, K., et al.: ToBI: a standard for labeling English prosody. In: Second International Conference on Spoken Language Processing (1992)

16. Tong, R., Ma, B., Zhu, D., Li, H., Chng, E.S.: Integrating acoustic, prosodic and phonotactic features for spoken language identification. In: Acoustics, Speech and Signal Processing, vol. 1, p. I. IEEE (2006)
17. Vaissière, J.: Language-independent prosodic features. In: Cutler, A., Ladd, D.R. (eds.) Prosody: Models and Measurements, pp. 53–66. Springer, Heidelberg (1983). https://doi.org/10.1007/978-3-642-69103-4_5