Prosodic Feature Criterion for Hebrew Using Different Feature Sets

Ben Fishman
Tel-Aviv University
Tel-Aviv, Israel
benf22@gmail.com

Irit Opher
Afeka Academic
College of Engineering
Tel-Aviv, Israel
irito@afeka.ac.il

***Abstract*—Prosody is essential for everyday human communication and provides important information about intention and meaning. It is used for subtle expressions such as sarcasm as well as for denoting more common expressions like questions or declarations and even can indicate the physiological or emotional condition of a speaker. In our previous work we presented a Prosodic Feature Criterion (PFC) for evaluating the prosodic nature of a feature that was extracted from speech signal. The PFC score provides us with a way to rank the features and determine whether an acoustic or spectral feature carries prosodic information. In this paper we continue to explore this mechanism, using the OpenSMILE toolkit, which is a standard set of features widely used for acoustic analysis and prosody research. Our experiments are carried out using a dataset of Hebrew utterances specifically designed for prosody research. We apply the PFC over each feature separately, thus ranking the different features. We then compare this ranking with classification based ranking of the same features. In addition we show visualization of the PFC idea using dimension reduction of multiple features representation. Both these tests, validate the use of the PFC score, for evaluating the prosodic nature of a feature in regards to specific prosody classes.**

***Index Terms*—Prosody, Prosodic Features, OpenSMILE, Hebrew Dataset**

I. Introduction

Prosody can be defined as the non-contextual information conveyed in speech utterances. Prosodic cues provide valuable information for human communication as prosody is used to express emotional states, attitudes and intentions [1]. It can also help in assessing mental or physiological states in some neurological diseases [2]. Prosody has been extensively researched in past years by linguists (e.g [3], [4], [5]) and by speech scientists e.g. in many speech based systems such as Text to Speech (TTS) [6], Speech Morphing [7] or Speech based Analysis [8]. Quite a lot of work has been done towards standardization of prosody annotation, for example ToBI [9] which is used for annotating tones and break indices or the IPrA Prosodic Alphabet [10]. Still, from an engineering point of view, little attention has been given so far to generalizing and formulating the acoustic features that represent the perceived prosodic building blocks. In [11] we presented our approach regarding such a formulation. We defined and evaluated an initial concept for grading the prosodic nature of a feature using two prosody classes. This concept can be extended to form a general framework that covers various scenarios (e.g. multiple prosody classes, a larger feature set or tonal languages). In this work we use a larger feature set, and apply our methodology to a subset of the openSMILE toolkit [12], that consists of a few thousand features, where many of them are considered to be related to prosody or emotions. OpenSMILE is widely used (e.g. [13], [14], [15]) and is commonly used for classification of emotions or prosodies. Reproducing our results for prosodic nature assessment using this feature set supports validation of our methodology.

II. Prosodic Feature Criterion (PFC)

In our previous work [11] we proposed a simple criterion for determining whether a specific feature conveys prosodic information, and to what degree. Let us consider a dataset with utterances spoken using a few prosody classes. We would like to test whether a feature F can be declared as prosodic, in the sense that it conveys information regarding all or some of the prosody classes in a dataset. The criterion we have defined is based on the following ideas: (1) A prosodic feature should be dependent on some prosodic manifestations (2) A prosodic feature should be independent of significant changes in ”other aspects” of the speech utterance when the same prosody is used. Currently, we limit the term “other aspects” to refer only to the utterance’s content. This means we expect to see little change in prosodic feature’s values when same or different content is expressed with the same prosody, while we expect to see a significant change in the feature’s values when different prosody classes are used, even if the content is the same.

Let us formulate the above ideas: suppose we have a set of utterances U_{pc}^{k}, where $p = 1, 2...N_P$ is an index representing the different prosodies in our dataset, $c = 1, 2, ...N_C$ represents the different content types, i.e. different phrases, and $k = 1, 2, ...K_{pc}$ runs through all utterances of type pc, i.e. all the utterances in our dataset with prosody p and phrase type c. We look at pairs of utterances that are either with the same prosody class or with different prosody classes and calculate the dissimilarity between all pairs of utterances for each feature F. We denote this feature as prosodic if the following requirements hold:

1) Requirement 1: the dissimilarity between the extracted features is sufficiently small, for most utterance pairs of

978-1-5386-6378-3/18/$31.00 ©2018 IEEE

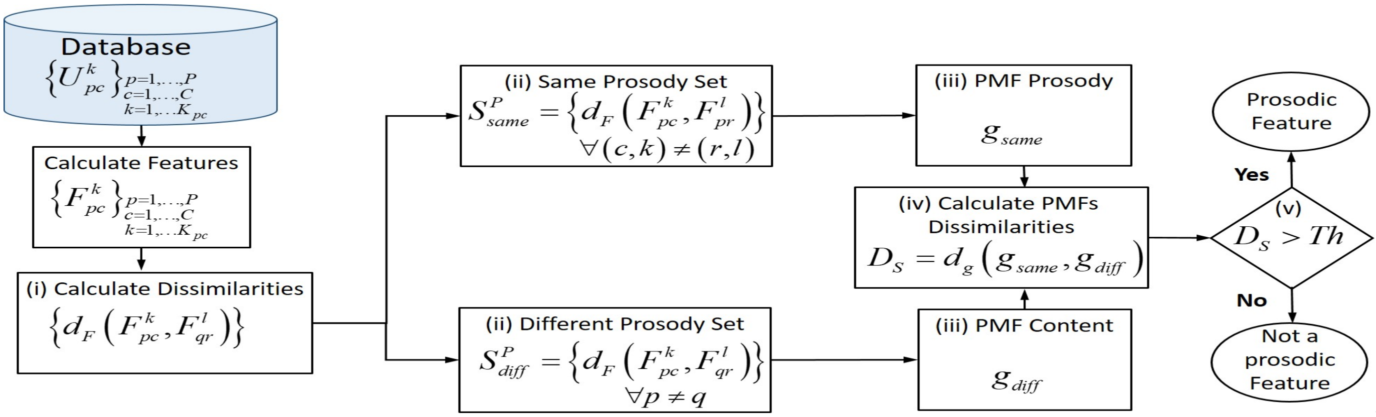
Database
$\{U^k_{pc}\}_{\substack{p=1,\dots,P\\c=1,\dots,C\\k=1,\dots K_{pc}}}$
Calculate Features
$\{F^k_{pc}\}_{\substack{p=1,\dots,P\\c=1,\dots,C\\k=1,\dots K_{pc}}}$
(i) Calculate Dissimilarities
$\{d_F(F^k_{pc},F^l_{qr})\}$
(ii) Same Prosody Set
$S^P_{same}=\{d_F(F^k_{pc},F^l_{pr})\}$
$\forall(c,k)\neq(r,l)$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

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

the same prosody p: $d_F\left(F^k_{pc},F^l_{pr}\right) < T_1$ 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, whether they are scalars or vectors.

2) *Requirement 2:* the dissimilarity between extracted features for utterance pairs of different prosody classes is higher than T_2 for most such pairs, $(q\neq p)$: $d_F\left(F^k_{pc},F^l_{qr}\right) > T_2$ for more than $x_2\%$ of the pairs.
These requirements should hold for both cases of same or different content type, but they are naturally stronger when the first requirement is applied under different content conditions and the second is applied under same content conditions.

T_1, T_2, x_1, x_2 can be tuned for each feature and for each dataset. To make this formulation simpler, we combine the two requirements in the following way -– we require that the Probability Mass Functions (PMFs) of the dissimilarities of the two examined sets (the set of utterance pairs of the same prosody and the set of utterance pairs of different prosody classes) will be well separated reflecting the different behavior of the feature's values for these two sets; This means that we require low values of dissimilarities for the same prosody set, and higher values of dissimilarity for the different prosody set. If, on the other hand, we find that there is a high similarity between the two PMFs (g_{same} and g_{diff}), we conclude that this feature does not carry any prosodic information, at least for the prosody classes that were used.

Based on these requirements we proposed a simple methodology to asses whether a single feature conveys prosodic information regarding some or all of the prosody classes in our data set. This methodology is described next, and is illustrated in Fig. 1 using a block diagram depicting the following steps: (i) Calculate dissimilarities between feature values over all possible utterance pairs (for simplicity we used the Euclidean distance as the dissimilarity function) (ii) Group all pairs into two sets - "same prosody" set (S^P_{same}) and "different prosody" set (S^P_{diff}) (iii) For each set, evaluate the Probability Mass Function (PMF) using the normalized histograms of the dissimilarities (g_{same} and g_{diff}) (iv) Calculate the dissimilarity score, denoted D_s, between the two PMFs (we used symmetrized KL-divergence as the dissimilarity function between g_{same} and g_{diff}) (v) Use a threshold Th to decide whether the feature can be considered prosodic, i.e. whether it conveys prosodic information.

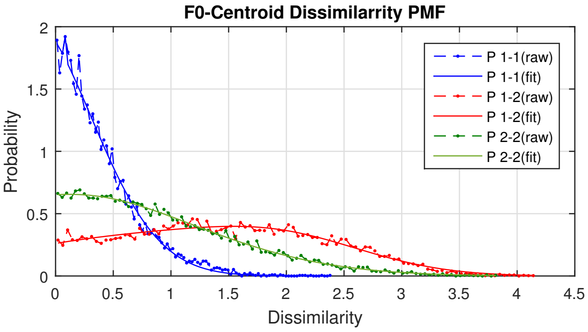

Figure 2. F0-Centroid dissimilarities curves. Same prosody (P1 vs. P1 in blue and P2 vs. P2 in green) and Different prosody (P1 vs. P2 in red)

III. PFC Evaluation

To evaluate our criterion, we use a special dataset and versatile feature sets:

A. Data Set

The dataset should be designed to allow us to check the two requirements defined in section II: it should consist of different prosodic classes, in order to check whether a feature represents prosodic manifestations for these classes. It should also include different content classes in order to verify that prosodic features are indeed independent of content changes. Hence, a suitable dataset requires every content class (i.e. every phrase) to be recorded in all prosody classes. Since finding this kind of a dataset is difficult, we collected our own dataset[1]. 36 speakers were recorded, both males and females of various ages. Overall there are 252 utterances in Hebrew.

[1] Hebrew dataset is available for research only by contacting the authors

Utterances are split into either three different content classes (i.e. phrases) or into two different prosody classes (question and neutral). Additional information can be found in [11].

B. Features

There are a few levels of features that can be extracted from a speech signal. First, the Lower Level Descriptors (LLD) that are usually calculated directly over the raw speech signal, and in most cases are evaluated separately for each frame or for each block of frames. Such features are F0, Energy, Voicing, etc'. The next level of features is called Functionals, that are calculated over the LLDs and constitute higher level descriptors, e.g. min, max, std and more. These functionals can be calculated using segments of various length, e.g frames, phonemes, syllables or utterances. Finally, feature values can be represented as a vector of the functionals values or as a scalar that represents yet a higher level of these functionals, e.g. std of F0 max values that were calculated over syllables.

In our framework, referring to a feature as "prosodic", means that it is related to some prosodic manifestations, definitely not to all of them. Hence, our evaluation is influenced by the choice of the dataset and the prosodies that are expressed in the data. In other words, if a feature gets a low PFC, it only means that the feature does not carry prosodic information regarding the prosody classes that were used in the dataset.

1) Initial Feature Set: To demonstrate and test our methodology, in [11] we used a 48 features set, most of which are considered standard in prosodic research, e.g. F0 and its derivatives, while some are hand-crafted features, such as duration-tilt and amplitude-tilt [8]. We also used some features that are not considered prosodic, e.g. MFCC, in order to show the difference between features that carry prosodic information and features that are related to other aspects of the speech signal.

2) Extended Feature Set - OpenSMILE: As mentioned above, we extended our feature set using the openSMILE toolkit [12] that is an open source toolkit for extracting many types of acoustic and spectral features. OpenSMILE can be used either off-line or on-line. This tool is widely used and has been cited over 1,300 times, mainly in the areas of speech recognition, emotion recognition, affective computing and music information retrieval. The openSMILE serves as a baseline acoustic feature set in many competitions, for example AVEC 2013 challenge [16] or at Interspeech, e.g. the 2009 emotion challenge [17], the 2010 paralinguistic challenge [18], the 2011 speaker state challenge [19], etc'. In this work We focus on the 2011 speaker state features set (*IS11_speaker_state.conf*). The challenge had two subtasks including the classification of "Alcohol Language" and "Sleepy Language". The feature set includes 4,368 features composed of LLDs (Energy, Spectra, voice related, etc.) and functionals applied over them. We chose this feature set configuration as it is large and widely used.

IV. Data Analysis

We use several methods to represent and compare the results to our initial work [11], to show the consistency of the conclusions, and validate our criterion:

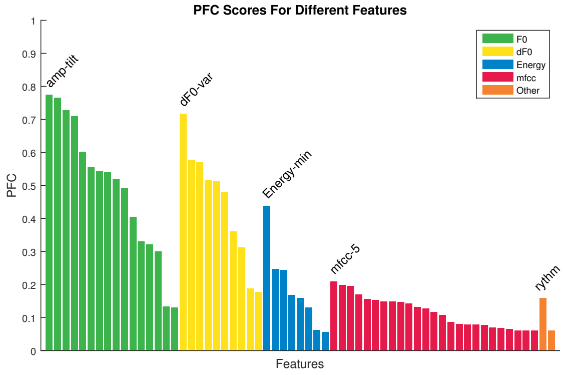

Figure 3. PFC scores for different feature families used in [11]; High scores for the F0 family (prosodic) and low scores for the non-prosodic MFCC family

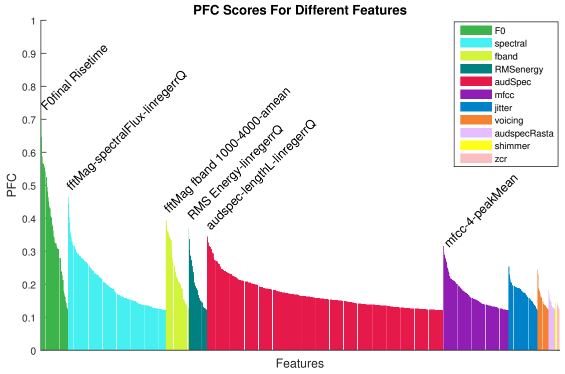

Figure 4. Comparison between features families. 1,000 best features out of OpenSMILE kit show similar results to Fig 3 as F0 receives higher scores than the Energy and MFCC families.

1) PFC of a Single Feature: As described in section II, in step (iii) of our methodology, we calculate the PMFs of the dissimilarity values, between pairs from same and different prosody groups. Fig. 2 shows an example for the F0-Centroid feature: instances with prosody P1 (neutral phrases) have low dissimilarity. Instances with P2 (question phrases) manifest higher values of dissimilarity probably due to the larger variance between different question types and expressions. Still, instances of P1-P2 pairs yield higher dissimilarity values, as expected for this feature that is definitely prosodic.

2) Analyzing PFC Scores by Features Families: We calculate PFC scores for a large number of features, separately for each feature, and group them to several features' families. In Fig 3 we see results over the feature set used in [11], that show high PFC scores for features of the F0 family, and low PFC scores for features of Energy and MFCC families. These results make sense as the MFCC family is known to be insensitive to prosody [20]. The Energy family is considered to be prosodic in nature [20], but due to the fact that this dataset includes only questions and neutral sentences, it makes sense that these features will receive low PFC scores. In Fig 4 we

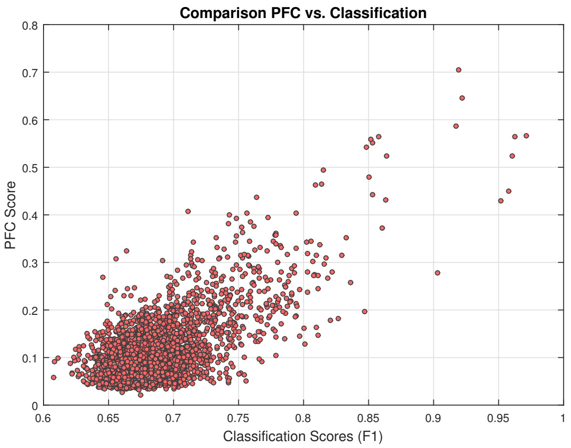

Figure 5. Comparison between PFC and classification F_1 results shows positive correlation. Both types of scores manifest similar tendency

see PFC scores for the best (i.e. highest PFC scores) 1,000 OpenSMILE features. We notice similar behavior, as the F0 family still receives the highest scores, while the Energy and MFCC families receive lower scores. An example of families that were not included in [11] are Jitter and Shimmer that are not dominant in differentiating between the prosody classes in our dataset, hence receive low PFC scores.

3) Comparison with Classification Results: PFC allows us to grade features by the amount of prosodic information they carry. We can then perform another grading process using standard methods used in classifications tasks, and compare the ranking of the features, obtained by the two sorting schemes.

As we relate only to a single feature's ranking, we compare our PFC-based ranking with ranking that is based on single feature classification results. For each feature, we trained a logistic regression classifier, while making sure train and test sets did not contain the same speakers. Then, we used a threshold that yielded the best F_1 measure[2] over the training set. Applying this threshold to the test set, we obtained classification accuracy for the test set. These accuracy scores were used to rank the different features for the classification task. Over the initial feature set, we compared the list of the best 15 features using PFC scores vs. a list of the best 15 features using F_1 scores and found out that 13 of the features are common. Further information can be found in [11]. Over the OpenSMILE features set, we found that out of the best 20 features using PFC or F_1 scores, 16 features were common. Next we illustrate correlation between the two ranking methods. In Fig 5 we see PFC scores vs. F1 scores for the OpenSMILE feature set. There is clearly a positive correlation between the two score types, and the correlation coefficient for this feature set is 0.68. It is important to clarify at this point that PFC is not a merely a feature selection method and conveys more information about features than does the F_1 measure. Hence, we expect only some positive correlation between these ranking methods.

4) Features' Visualization: Another way of demonstrating that a specific feature carries prosodic information is to look at its values for different classes, and see whether they are separable in some space. In 1D space (i.e. a single scalar feature) we can look at the PMFs of the feature values. In Fig. 8 we see an example of a feature that separates well between two prosodic classes and does not separate at all the different content classes, thus it can be considered a prosodic feature in regard with the prosody classes under investigation.

When using more features, we need to visualize high-dimensional data in order to check separability between the different classes. This can be done by applying dimension reduction schemes. In [11] we chose the t-SNE algorithm [21], and applied it over the best PFC features out of the initial feature set and showed very good separation between two prosodic classes. There was no separation at all between

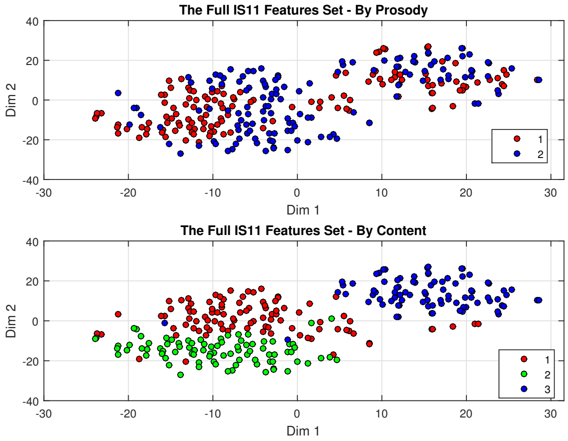

Figure 6. Dimension reduction of the full Interspeech 2011 features set including 4,368 features, showing better separation for content classes

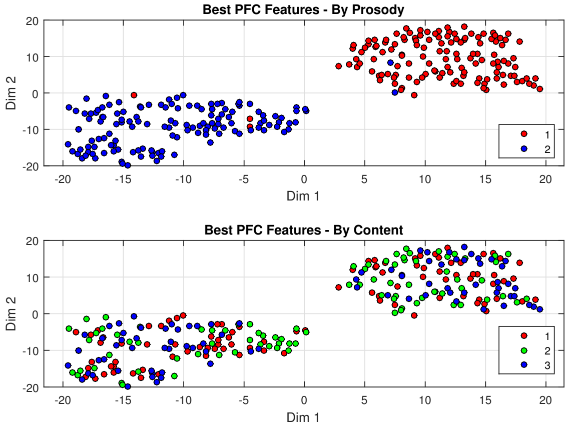

Figure 7. Dimension reduction of the best 14 PFC features, showing better separation for prosody classes

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

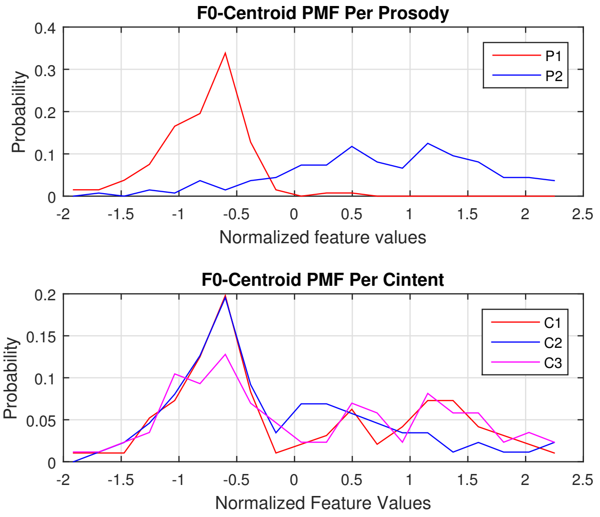

Figure 8. F0-Centroid PMF by prosody classes (top) and by content classes (bottom), showing this feature can be considered prosodic

the content classes. In this work, we apply t-SNE over: (1) the full 4,368 features space, and (2) the best 14 prosodic features obtained using the PFC. Fig. 6 shows that the whole feature set did not achieve good separation between prosody classes at all, and actually provided good separation between the three content classes. Fig. 7 shows the opposite: dimension reduction over a subset of the best PFC features, shows clear separation between the prosody classes. This means that even though this feature set is actually biased towards content representation, the PFC succeeded in pinpointing a minimal set of features that indeed carries prosodic information.

V. Conclusions and Future Work

In this paper we have shown that the PFC score presented at [11] can be applied to a large standard feature set and provide important information regarding the prosodic nature of some of the features in the OpenSMILE set. The relevance of this score was validated by: (1) Comparing the features ranking induced by the PFC score to a ranking calculated according to the performance of a single feature classification task (where we classify utterances into two prosody classes). (2) Dimension reduction algorithm that is used for visualization of multiple features. We showed that even though the full OpenSMILE set is biased towards content classes separation, we were able to find, using the PFC score, some features that carry prosodic information. Future work will address the case of more than two prosody classes, additional experiments with different datasets and different languages and exploration of more elaborated schemes including a mathematical formulation of the representation of prosodic information.

References

[1] O. Pierre-Yves, “The production and recognition of emotions in speech: features and algorithms,” *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 157–183, 2003.

[2] J. J. Diehl and R. Paul, “The assessment and treatment of prosodic disorders and neurological theories of prosody,” *International journal of speech-language pathology*, vol. 11, no. 4, pp. 287–292, 2009.

[3] A. Wennerstrom, *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press, 2001.

[4] P. Trofimovich and W. Baker, “Learning second language suprasegmentals: Effect of l2 experience on prosody and fluency characteristics of l2 speech,” *Studies in second language acquisition*, vol. 28, no. 1, pp. 1–30, 2006.

[5] P. K. Kuhl, “Early language acquisition: cracking the speech code,” *Nature reviews neuroscience*, vol. 5, no. 11, p. 831, 2004.

[6] S.-H. Chen, S.-H. Hwang, and Y.-R. Wang, “An rnn-based prosodic information synthesizer for mandarin text-to-speech,” *IEEE transactions on speech and audio processing*, vol. 6, no. 3, pp. 226–239, 1998.

[7] A. Qavi, S. A. Khan, and K. Basir, “Voice morphing based on spectral features and prosodic modification,” in *Multi-Topic Conference (INMIC)*. IEEE, 2014, pp. 401–405.

[8] L. Mary and B. Yegnanarayana, “Extraction and representation of prosodic features for language and speaker recognition,” *Speech communication*, vol. 50, no. 10, pp. 782–796, 2008.

[9] 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 *Second international conference on spoken language processing*, 1992.

[10] J. Hualde and P. Prieto, “Towards an international prosodic alphabet (ipra),” *Laboratory Phonology*, vol. 7, 1 2016.

[11] B. Fishman, I. Lapidot, and I. Opher, “Prosodic features’ criterion for hebrew,” in *proceedings of International Conference on Text, Speech, and Dialogue*. Springer, 2018.

[12] Eyben, Wöllmer, and Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[13] M. Man-Wai, “Feature selection and nuisance attribute projection for speech emotion recognition.”

[14] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, “Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 275–278.

[15] M. Tahon and L. Devillers, “Towards a small set of robust acoustic features for emotion recognition: challenges,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 1, pp. 16–28, 2016.

[16] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “Avec 2013: the continuous audio/visual emotion and depression recognition challenge,” in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[17] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[18] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, “The interspeech 2010 paralinguistic challenge,” in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[19] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, “The interspeech 2011 speaker state challenge,” in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[20] N. Singh, R. Khan, and R. Shree, “Mfcc and prosodic feature extraction techniques: A comparative study,” *International Journal of Computer Applications*, vol. 54, no. 1, 2012.

[21] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, pp. 2579–2605, 2008.