INTERPRETABLE PHONOLOGICAL FEATURES FOR CLINICAL APPLICATIONS

Yishan Jiao[1]*, Visar Berisha*[1,2] *and Julie Liss*[1]

[1]Department of Speech and Hearing Science
[2] School of Electrical, Computer, and Energy Engineering
Arizona State University

ABSTRACT

Instrumental analysis of speech sometimes complements subjective evaluations in speech and language therapy; however, apart from elemental speech features such as pitch and formant statistics, higher dimensional spectral features are rarely used in practice because they are clinically uninterpretable. While these features are likely to somehow be related to clinical intervention, this relationship remains to be determined. This paper uses artificial recurrent neural networks to map high-dimensional spectral features into phonological features that are easily interpretable and provide fine-resolution information regarding articulation quality. The evaluation on a dysarthric speech data set shows strong correlation between the phonological feature measures and perceptual ratings. To increase clinical utility, we provide a new way to visualize phonological disturbances that provides clinicians with actionable information about intervention strategies.

Index Terms— phonological features, recurrent neural networks, clinical applications

1. INTRODUCTION

In speech signal processing, there is no shortage of acoustic features that can be extracted for various tasks. For example, the mel-frequency cepstrum coefficients (MFCCs) are widely used in automatic speech recognition [1][2]; the linear prediction coefficients (LPCs) are well-studied in speech coding [3][4]; the line spectral frequencies (LSFs) or the line spectral pairs (LSPs) are commonly used in text-to-speech synthesis and voice conversion [5][6]; the perceptual linear predictive (PLP) coefficients have shown outstanding performance in speaker identification [7][8]. Recently with the development of deep learning, the output or the intermediate output of artificial neural networks can also serve as acoustic features for certain tasks [9][10][11]. However, in the clinical practice of speech therapy, only a subset of basic lower dimensional features are used (e.g., fundamental frequency, formant frequencies, jitter and shimmer) [12][13]. Other more complex features, including the ubiquitous MFCCs, are rarely used by clinicians because of their lack of interpretability. However, these high dimensional features contain a great deal of information about the patient and the disease. In fact, a number of engineering studies of pathological speech analysis use these acoustic features to automatically make predictions about the disease state [14][15][16][17]; however, most of these methods only provide the final prediction results to clinicians without attempting to make explain why the decision was made. Rather than automatically making the decision for the clinician, we posit that it makes more sense to provide more interpretable features that allows the clinicians to make better decision themselves.

Phonological features [18], namely class, manner and place of speech sounds, are more comprehensible to clinicians than the traditional high-dimension acoustic features often used in speech analysis applications. Chomsky and Halle developed a phonological feature system called the 'Sound Pattern of English' (SPE) [19]. In this system, each phone can be represented by a vector of binary values that corresponds to a comprehensive set of production features including, sonorant, high/low (tongue position during vowel), round (lip rounding), etc. These dimensions of production are well understood by clinicians as they are an integral part of all speech science courses. If we can automatically evaluate the speech based on these specific phonological categories, clinicians will be able to infer a great deal of information about the articulation abilities of a patient. Therefore, in this paper, we attempt to map the difficult-to-interpret acoustic features onto understandable phonological features and propose an estimate of articulation quality based on these features.

A few other studies have explored the relationship between traditional acoustic features and phonological features [20][21][22]. In [20], King and Taylor proposed to use recurrent neural networks (RNNs) to detect phonological features in continuous speech. In this paper, we follow their study and use a multi-label RNN to map acoustic features to phonological features. The difference between our study and [20] are as follows: (1) the purpose of our study is to use phonological features to evaluate the articulation quality of dysarthric speakers instead of simply inferring phonological features for healthy speech; (2) The model used in [20] was actually a neural network with time-delayed recurrent connections, while in this paper, we use long short-term memory (LSTM) RNNs with multiple layers so that the

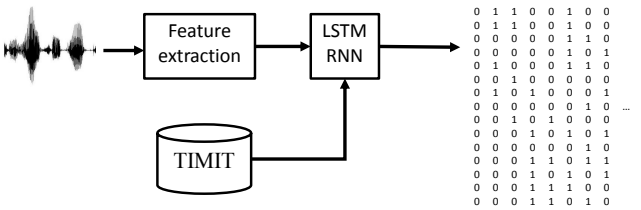

Fig. 1. Diagram of the classification system.

network learns longer time dependencies; (3) To estimate the articulation quality based on these features, we train RNNs on healthy speech and apply the model on pathological speech so that the results reveal atypical speaker-specific articulation patterns; (4) We propose a new visualization tool geared to clinical applications. In the area of pathological speech analysis, Wong et al also attempted to explore the articulatory characteristics of dysarthric speech using similar phonological features which they call 'distinctive features' [23][24]. However, the analysis in their study was based on single words with hand labeled phonetic segmentation. In contrast, all analyses on dysarthric data in our study are based on continuous speech without any manual segmentation.

The organization of this paper is as follows. Section 2 introduces the implementation of the phonological feature detection algorithm and the measure of articulation quality. The evaluation on dysarthric speech is described in Section 3. We make concluding remarks in Section 4.

2. PHONOLOGICAL FEATURE DETECTION AND REPRESENTATION

2.1. Building RNNs on healthy speech

Recurrent neural networks are state-of-the-art machine learning models that have recently been used in many speech signal processing areas, such as ASR [10], speech synthesis [25], speech enhancement [26], speaking rate estimation [27], accent identification [28], etc. The main advantage of RNNs over traditional neural networks is that they can learn long-term dependencies in multi-dimensional time-series sequences (e.g., features extracted periodically from speech signal). Therefore, the prediction that the RNN makes for a particular speech frame depends not only on the features of the current frame but also on the features preceding and/or following it. This is consistent with speech production, where articulation is dependent not only on the current phoneme being produced but also on the ones preceding and following it (co-articulation). Therefore, in this paper we propose to use RNN for phonological feature classification by building a mapping from MFCCs to SPE features.

Fig. 1 shows a block diagram of the proposed system. We use the TIMIT database to train the system by mapping acoustic features extracted from each frame of speech to SPE

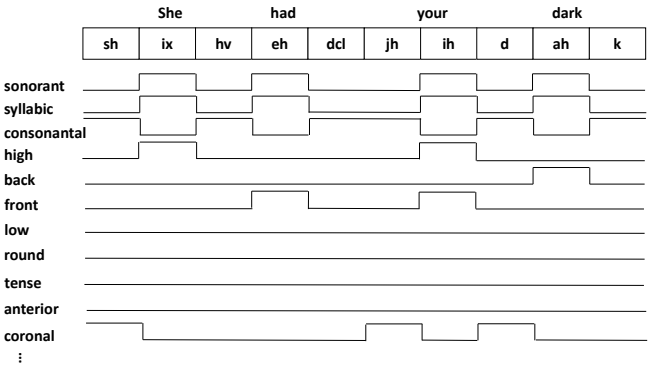

Fig. 2. From phoneme to phonological features.

labels. After training, the model can automatically generate the binary SPE labels for an input speech sample. We describe the system in detail below.

The model was trained on the classic TIMIT healthy speech database [29], which contains 630 speakers of eight major dialects of American English and each speaker has 10 speech samples sampled at 16-bit, 16kHz with phonetic labels. The training set contains 462 speakers (4620 samples) and the remaining (168 speakers, 1680 samples) are in the test set. The original SPE system contains 22 feature classes which are believed sufficient for analyzing the phonemes of any language. For English, only 13 feature classes are needed and the remaining are regarded as redundant [19]. Brøndsted expanded the 13 features into 15 and set rules to map each TIMIT phoneme into these features [30]. Please refer to [30] for the specific mapping between SPE features and TIMIT phonemes. We followed his rules and transformed all the phonemes in TIMIT using these 15 SPE features: sonorant, syllabic, consonantal, high, back, front, low, round, tense, anterior, coronal, voice, continuant, nasal, and strident. Fig. 2 shows an example of this transformation for the first part of a popular TIMIT utterance.

Table 1. The classification accuracy of RNN for each phonological feature on the TIMIT test set.

Phonological Feature	Accuracy (%)	Phonological Feature	Accuracy (%)
sonorant	96.14	syllabic	92.09
consonantal	91.47	high	89.68
back	94.52	front	94.04
low	93.71	round	93.93
tense	97.43	anterior	92.02
coronal	90.86	voice	91.74
continuant	93.61	nasal	97.99
strident	97.52		

The speech samples were analyzed using a 20ms Hamming window with 10ms overlap. 13th-order MFCCs and

Table 2. Correlation coefficients between phonological features with perceptual ratings.

sonorant	-0.30	**syllabic**	**-0.43**
consonantal	**-0.71**	**high**	**-0.42**
back	0.17	front	-0.18
low	0.08	round	0.11
tense	-0.16	**anterior**	**-0.69**
coronal	**-0.67**	**voice**	**-0.63**
continuant	**-0.59**	nasal	-0.10
strident	**-0.69**	**Linear Combination**	**0.79**

delta and delta-delta coefficients were extracted from each frame. All frames in a segmented phoneme were labeled with a vector of 15 binary values, leading to a multi-label classification problem. The architecture of the RNN is as follows. It has an input layer with 39 nodes (input feature dimension); three hidden layers, with 156, 256, 156 bidirectional long short-term memory (LSTM) nodes; and an output layer with 15 softmax nodes. The training objective was to minimize the cross entropy-error between the distribution of the neural network output and the ground-truth. The open source toolbox CURRENNT [31] was used to train the network. The RNN was trained with the TIMIT training set and evaluated on the TIMIT test set. The classification accuracy of the model in each feature dimension on the test set is shown in Table 1. From the table, we can see the accuracies for all features are close to or over 90%.

2.2. Assessing the articulation abilities of dysarthric speakers using phonological features

The RNN learns a 'healthy' model for each of the phonological features. We propose to use this healthy model to assess articulation in dysarthric speakers. Since the model is trained on healthy speakers and we apply it to dysarthric speakers, we can use the distance between the dysarthric speaker and healthy speakers in the phonological feature space to assess articulation. For example, consider a dysarthric speaker with hypo-nasality. In this example, the nasal phonemes the patient produces are likely to be de-nasalized. Therefore, the number of frames detected as 'nasal' should be fewer than for a healthy speaker when they speak the same content at the same rate.

To find an articulatory distance between one dysarthric speaker and a set of healthy speakers, we first transform the raw phonological extracted features into a set of proportions. Suppose the total number of frames in a speech sample provide by speaker i is N_i. Among them, N_i^p frames were detected as belonging to the p^{th} phonological feature class. We convert all detected phonological features to percentages by normalizing with the total number of frames: $\frac{N_i^p}{N_i}$, where $i = 1, 2, ..., S$, $p = 1, 2, ..., 15$, and S is the number of speakers. If we have a database of healthy speech, we can estimate a healthy distribution for each of the phonological feature percentages. This allows us to estimate how far a dysarthric speaker is from the healthy distribution for a specific feature by using the Mahalanobis distance (MD) [32] which measures the distance between a data point (dysarthric speaker) and a distribution (a group of healthy speakers). Suppose the estimate of the p^{th} phonological feature of the dysarthric speaker j is $e_j^p = \frac{N_j^p}{N_j}$, its distance from the healthy distribution for that feature is given by

$$D(e_j^p) = \sqrt{\frac{(e_j^p - \hat{\mu_p})^2}{\hat{\sigma_p}^2}} \quad (1)$$

where $\hat{\mu_p}$ and $\hat{\sigma_p}$ are the estimated mean and standard deviation of healthy group's distribution for feature p.

3. EVALUATION ON DYSARTHRIC SPEECH

3.1. Database

We evaluate the proposed system using a dysarthric speech database collected at the Motor Speech Disorders Lab at Arizona State University [33]. There are 33 speakers in the dataset with dysarthria subtypes as follows: ataxic dysarthria secondary to cerebellar degeneration (n = 11), mixed flaccid-spastic dysarthria secondary to amyotrophic lateral sclerosis (n = 10), hyperkinetic dysarthria secondary to Huntington's disease (n = 4), hypokinetic dysarthria secondary to Parkinson's disease (n = 8). Along with the dysarthric speech samples, there were another 13 healthy speakers recorded in parallel, which means they read the same content as the dysarthric speakers. The speech was sampled at 44.1 kHz with 16 bit resolution. Each speaker was recorded when reading 5 phoneme-balanced sentences. Six speech language pathologists (SLPs) were asked to listen to these speech samples and rate the articulatory precision for each speaker. The rating was on a 7-point scale (1 = normal, 7 = severe deviation from normal). The Evaluator Weighted Estimator (EWE) was used to combine the multiple ratings into a single one by calculating the mean value weighted by individual reliability [34]. The details of the speech database can be found in [33].

3.2. Correlation analysis

In our experiment, the speech was first down sampled to 16kHz and segmented into frames. Features were extracted as described in Section 2.1. The RNN trained on TIMIT was applied to the acoustic features from dysarthric speakers. For every speech segment, the model provided a 15-dimensional binary output that predicts which of the phonological feature classes the frame belongs to. The ones in the binary sequence

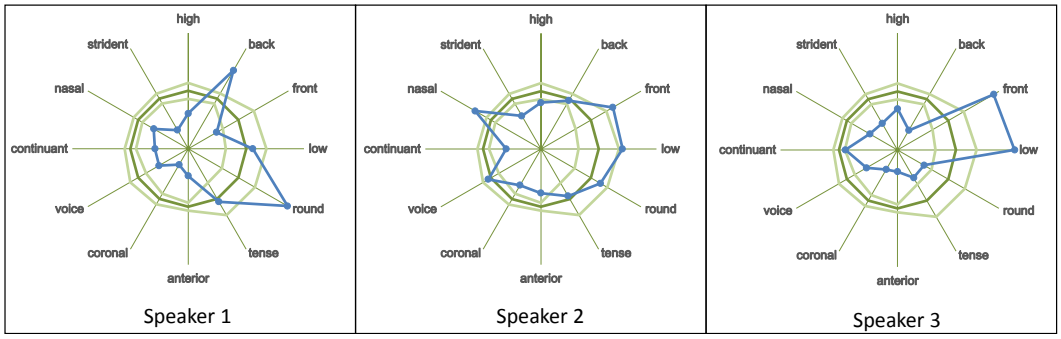

Fig. 3. Visualization of phonological features in a spider plot.

indicate that the model has detected a target feature, while zeros indicate that the feature was not detected for that particular frame.

For each speaker, we concatenated the results of five sentences together and calculated the proportion of detected frames in the sequence for each phonological category and calculated its MD to the healthy group per Eqn. 1. The Pearson correlation between the estimated distances with the EWE subjective rating was calculated. The result is shown in Table 2. The features with absolute correlation coefficient values higher than 0.4 are highlighted in the table. From the table, we can see that some of the features such as consonantal, anterior, coronal, voice, and strident, have strong correlation ($[0.60, 0.79]$) with the perceptual rating; others such as syllabic, high, and continuant have moderate correlations ($[0.40, 0.59]$) with the perceptual rating. A linear regression on the highlighted phonological features was conducted to predict the perceptual rating. The leave-one-out cross validation results are shown as the last item of Table 2, and it shows a strong correlation.

3.3. Visualization for clinical applications

A key feature for the phonological features we propose is interpretability. To that end, we propose a new visualization tool based on the spider plot to integrate all phonological features into a single representation (see Fig. 3) Since the first three features (sonorant, syllabic and consonantal) are a collection of all other features, we did not show them in this plot. The average value of the estimated phonological features from healthy group was scaled to one, corresponding to the unit circle (dark green lines) in the figure. The standard deviation for each feature is shown as light green lines. The values of the evaluated dysarthric speaker, normalized by the average of healthy speakers ($\frac{e_j^p}{\mu_p}$) is shown as the blue lines in the figure. For each phonological group, the closer the speaker is to the unit circle, the more precise his or her ability to produce sounds from that category. This representation provides clinicians with an informative impression of the articulation abilities of the speaker. For example, in Fig. 3, we show three dysarthric speakers with different profiles. We can see from the plot that Speaker 1 produced more round and back vowels while fewer high and front vowels than healthy speakers; furthermore, the speaker produced fewer clear consonant. Speaker 2 seems to have less severe articulation problem than the other two since the figure shows little deviation from the healthy distribution. Speaker 3 shows imprecise vowel and consonant production; however with a different distribution than speaker 1. This representation provides actionable information to clinicians that seek to improve various aspects of articulation through intervention.

4. CONCLUSIONS

High dimensional acoustic features are usually hard to interpret in clinical practice. Phonological features, namely class, manner and place of speech sounds make more sense to clinicians. This paper has proposed to use the LSTM RNN to map acoustic features to SPE binary phonological features. The model was trained on healthy speech and applied on dysarthric speech so that the results could reveal the specific pathological articulation patterns of a speaker. The model compares the articulatory abilities of a speaker against a group of healthy controls using a fixed read passage. Evaluation on a dysarthric speech database revealed that most of the articulatory features are moderately or strongly correlated with perceptual impressions of articulatory precision. To make the result more interpretable, a spider plot has been developed for visualization. From the plot, clinicians can gain valuable information regarding the articulation abilities of a speaker, which can be helpful for developing personal treatment plans and monitoring disease progress.

5. ACNOWLEDGEMENT

This work was partially supported by an NIH 1R21DC013812 grant. The authors graciously acknowledge a hardware donation from NVIDIA.

6. REFERENCES

[1] Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of speech recognition," 1993.

[2] Jean-Claude Junqua and Jean-Paul Haton, *Robustness in automatic speech recognition: Fundamentals and applications*, vol. 341, Springer Science & Business Media, 2012.

[3] Alan V McCree and Thomas P Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[4] B Atal and J Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82*. IEEE, 1982, vol. 7, pp. 614–617.

[5] Paul Taylor, *Text-to-speech synthesis*, Cambridge university press, 2009.

[6] Levent M Arslan and David Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum.," in *Eurospeech*, 1997.

[7] Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[8] Jiahong Yuan and Mark Liberman, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3878, 2008.

[9] Navdeep Jaitly and Geoffrey Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5884–5887.

[10] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[11] Karel Veselỳ, Martin Karafiát, and František Grézl, "Convolutive bottleneck network features for LVCSR," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 42–47.

[12] Joseph R Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*, Elsevier Health Sciences, 2013.

[13] Ray D Kent and Y-J Kim, "Toward an acoustic typology of motor speech disorders," *Clinical linguistics & phonetics*, vol. 17, no. 6, pp. 427–445, 2003.

[14] Alireza A Dibazar, S Narayanan, and Theodore W Berger, "Feature analysis for automatic detection of pathological speech," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*. IEEE, 2002, vol. 1, pp. 182–183.

[15] Catherine Middag, Jean-Pierre Martens, Gwen Van Nuffelen, and Marc De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–9, 2009.

[16] Karthikeyan Umapathy and Sridhar Krishnan, "Feature analysis of pathological speech signals using local discriminant bases technique," *Medical and Biological Engineering and Computing*, vol. 43, no. 4, pp. 457–464, 2005.

[17] Alan Wisler, Visar Berisha, Karthikeyan Ramamurthy, Andreas Spanias, and Julie Liss, "Removing data with noisy responses in regression analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2066–2070.

[18] Roman Jakobson, Gunnar Fant, and Morris Halle, "Preliminaries to speech analysis. The distinctive features and their correlates," 1951.

[19] Noam Chomsky and Morris Halle, "The sound pattern of English.," 1968.

[20] Simon King and Paul Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 333–353, 2000.

[21] Afsaneh Asaei, Milos Cernak, and Hervé Bourlard, "On compressibility of neural network phonological features for low bit rate speech coding," in *Proc. of Interspeech*, 2015, pp. 418–422.

[22] Milos Cernak, Blaise Potard, and Philip N Garner, "Phonological vocoding using artificial neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4844–4848.

[23] Ka Ho Wong, Yu Ting Yeung, Patrick CM Wong, Gina-Anne Levow, and H Meng, "Analysis of dysarthric speech using distinctive feature recognition," in *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015, p. 86.

[24] Ka Ho Wong, Wing Sum Yeung, Yu Ting Yeung, and Helen Meng, "Exploring articulatory characteristics of Cantonese dysarthric speech using distinctive features," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6495–6499.

[25] Heiga Zen and Haşim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4470–4474.

[26] Felix Weninger, Florian Eyben, and Björn Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3709–3713.

[27] Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss, "Online speaking rate estimation using recurrent neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5245–5249.

[28] Yishan Jiao, Ming Tu, Visar Berisa, and Julie Liss, "Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features.," in *Interspeech*, 2016.

[29] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic data consortium, Philadelphia*, vol. 33, 1993.

[30] Tom Brøndsted, "A SPE based distinctive feature composition of the CMU label set in the TIMIT database," *Report of speech research activities. Center for PersonKommunikation, Aalborg University*, 1998.

[31] Felix Weninger, Johannes Bergmann, and Björn Schuller, "Introducing CURRENNT–the munich open-source CUDA RecurREnt neural network toolkit," *Journal of Machine Learning Research*, vol. 16, no. 3, pp. 547–551, 2015.

[32] Prasanta Chandra Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences (Calcutta)*, vol. 2, pp. 49–55, 1936.

[33] Julie M Liss, Laurence White, Sven L Mattys, Kaitlin Lansford, Andrew J Lotto, Stephanie M Spitzer, and John N Caviness, "Quantifying speech rhythm abnormalities in the dysarthrias," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1334–1352, 2009.

[34] Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.