EVALUATION OF SYLLABLE RATE ESTIMATION IN EXPRESSIVE SPEECH AND ITS CONTRIBUTION TO EMOTION RECOGNITION

Mohammed Abdelwahab and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
Email: mxa129730@utdallas.edu, busso@utdallas.edu

ABSTRACT

It is commonly accepted that speaking rate is an important aspect characterizing expressive speech. The speaking rate increases for emotions such as happiness and anger, and decreases for emotions such as sadness. In spite of these observations, most of the current speech emotion classifiers do not explicitly use speaking rate features. This study explores two interrelated questions to evaluate the role of speaking rate in emotion recognition: Can we reliably estimate syllable rate from emotional speech? Does syllable rate provide complementary emotional information over other acoustic features? We consider two syllable rate estimation algorithms, as well as reference values derived from forced alignment. We evaluate the performance of these syllable rate estimation methods in expressive speech (SEMAINE database). The analysis reveals a drop in performance as the intensity of the emotion increases. Next, we conduct emotion recognition experiments to evaluate the contribution of syllable rate in recognizing emotions. The emotion classification experiments demonstrate that features conveying accurate syllable rate estimations complement features that are commonly used in current emotion recognition system.

Index Terms— Speech rate, speech emotion recognition, prosody and emotion

1. INTRODUCTION

Emotion plays an integral part of human communication, affecting our rational and intelligent decisions, and the manner in which we interact with others [1]. Therefore, modeling and detecting expressive behaviors in speech has emerged as an important research area to improve *human-computer interaction* (HCI). A robust emotion recognition system can benefit applications in the area of health informatics, education, and entertainment [2]. A key problem in recognizing emotions from speech is identifying emotionally discriminative acoustic features [3, 4]. This study focuses on the role of speech rate in emotion recognition.

Speech rate conveys important information that can be exploited in speech processing tasks. It has been used as feature to characterize fluency in a language [5], and cognitive load on the speaker [6]. Speaking rate affects how a person is perceived (e.g., personality traits). For example, people whose speaking rate is higher than normal are perceived as more competent [7]. People speaking too slow or too fast are perceived as less benevolent and truthful, showing an inverted U-relationship with speech rate [8]. Speech rate also affect the performance of *automatic speech recognition* (ASR) systems, where speech rate variations degrade recognition performance due to mismatches between train and test conditions [9–11]. In emotion recognition, speech rate is usually described as an appealing discriminative feature [12, 13]. Studies have consistently reported that speaking rate increases for emotions such as happiness and anger, and decreases for emotions such as sadness and boredom [4]. Some of the commonly used *duration* features include the ratio between voiced and unvoiced segments [14–17], zero crossing rate [18]. While these features may be correlated with speech rate, current emotion recognition systems usually do not include any direct metric of speech rate.

Recent advances on speech processing have resulted on toolkits to directly estimate syllable rate. This study considers the algorithms proposed by Wang and Narayanan [19] and de Jong and Wempe [20]. An open question is whether these algorithms can robustly estimate speech rate metrics for expressive speech. The features derived from emotional speech differ from the acoustic properties of neutral, nonemotional speech [21]. Given that syllable rate algorithms are trained with neutral speech, it is not clear how robust these algorithms are for expressive speech. This study addresses this question by quantifying the performance of these syllable rate systems for sentences with different values of valence (negative versus positive) and activation (calm versus active). The evaluation reveals a drop in performance as the activation and valence values increase. Our second related question is whether these metrics provide additional emotion information over the features that are currently used in speech emotion recognition systems. We address this question by conducting emotion

This work was funded by NSF (IIS-1217104 and IIS-1329659).

recognition evaluations, where the classifiers are separately trained with selected group of features (voice quality, spectral, RASTA, MFCC, F0 and energy features). We quantify the improvement or drop in accuracy observed when the syllable rate features are included in addition to other feature sets. When these features are accurately estimated, syllable rate metrics provide complementary information, especially for spectral features.

2. RESOURCES

2.1. Syllable Rate Estimation

Studies have investigated automatic algorithms to estimate syllable rate given the role of speech rate on human communication. Faltlhauser et al. [22] used *Gaussian mixture models* (GMM) to detect slow, medium and fast speech rate. The likelihood scores of the three classes were used as input of a neural net, which provided a continuous, online estimate of speech rate. Zhang and Glass [23] applied sinusoid fitting on energy peaks to predict possible regions where syllable nuclei can appear. Next, a simple slope based peak counting algorithm was used to get the positions of the syllable nuclei. Yuan and Liberman [24] used a broad phonetic class recognizer for syllable detection and speech rate estimation.

Morgan et al. [25] presented a syllable rate estimator called *enrate*, which considers the first spectral moment of the broad-band energy envelope. The results from *enrate* were correlated with the transcribed syllabic rate, but the deviations were large. They later proposed *mrate*, which combines the *enrate* approach with point-wise correlation between pairs of compressed sub-band energy envelopes [26]. Wang and Narayanan [19] extended this approach by considering only prominent subbands. In addition to spectral correlation, they added temporal correlation and other strategies to improve robustness against neighboring syllable smearing and spurious syllable envelope peaks. De Jong and Wempe [20] wrote a script for Praat [27] that detects syllables nuclei by peaks in intensity. It counts the number of intensity peaks, with drops in intensity of at least 2 dB immediately before and after the peak. This approach is conducted on voiced speech, determined by the F0 contour.

The present study uses the algorithms proposed by Wang and Narayanan [19] and de Jong and Wempe [20] to automatically estimate syllable rate. To the best of our knowledge, this is the first study that evaluates the performance of speech rate algorithms on emotional speech, and explores whether speech rate contains complementary information to the common feature sets used in speech emotion recognition systems.

2.2. SEMAINE Database

The study relies on the *SEMAINE* database [28], which is an audiovisual database with natural emotional displays. The corpus includes sessions recorded from two individuals, an *operator* and a *user*, interacting through teleprompter screens from two different rooms. The emotions were elicited with the *sensitive artificial listener* (SAL) framework, where the operator assumes four personalities aiming to elicit positive and negative emotional reactions from the user. The data was recorded using five high resolution, high frame-rate cameras, and four microphones. This study uses emotional data from 24 speakers (users) interacting with the operators. The transcriptions of the dialogs are available. In total, we consider 2830 turns, discarding segments with duration less than 0.5s.

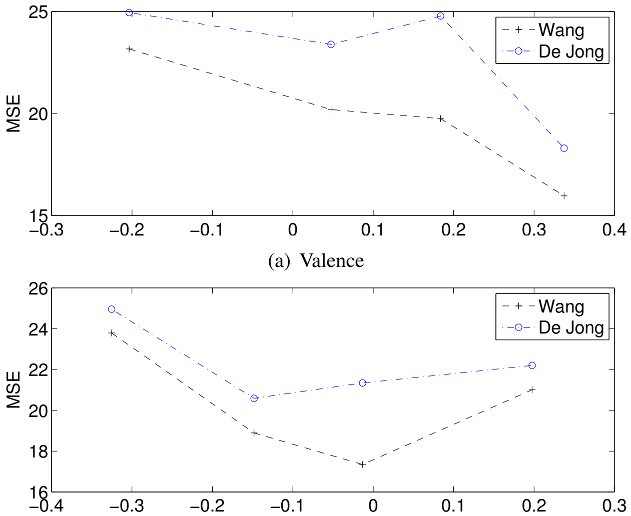
The sessions were emotionally annotated by 6-8 raters. Instead of assigning global labels to the speaking turns, evaluators provided time-continuous emotional traces using the FEELTRACE toolkit [29]. As the evaluators watch the recordings, they move the mouse cursor over a *graphical user interface* (GUI), where the axes represent specific emotional attributes. The interface records the position of the cursor, providing a continuous profile, or trace, for that emotional dimension. Among other descriptors, the perceptual evaluation considered the dimensions activation (calm versus active), valence (negative versus positive), control (weak versus strong) and expectation (predictable versus unexpected) [30]. This study focuses on activation and valence, which are the most commonly used emotional dimensions.

3. SYLLABLE RATE ESTIMATION ANALYSIS

We first study the accuracy of current speech rate estimation algorithms when tested on emotional speech. These algorithms are trained with emotionally neutral speech, so we expect a drop in performance due to the mismatch between train and test conditions. In particular, we are interested in evaluating the performance of the syllable rate algorithms developed by Wang and Narayanan [19] and de Jong and Wempe [20] for sentences that are perceived with different values of valence and activation.

As mentioned in Section 2.2, the SEMAINE corpus was evaluated with time-continuous emotional traces using FEELTRACE. We split the corpus into speaking turns. Next, we derive global scores per turn for valence and activation using the traces provided by multiple evaluators. The approach consists in estimating the average value of the trace across different evaluators over the duration of the turn. After estimating the scores for valence and activation for each of the turn, we sort the speaking turns according to their values. For each emotional dimension, we separate the turns into four groups with equal sizes by estimating the first, second and third quantile over the scores. We will use these groups to evaluate the accuracy of syllable rate estimation for emotional turns perceived with different values of valence and activation.

We use forced alignment to define reference values for the syllable rate of the SEMAINE turns. We use these reference values to evaluate the performance of the speech rate algorithms. First, we obtain the phonetic alignment using



(a) Valence

(b) Activation

Fig. 1. Median values of MSE for the syllable rate estimation algorithms evaluated with emotional speech. The study considers the algorithms presented by Wang and Narayanan [19] and de Jong and Wempe [20]. The figure shows MSE versus (a) valence, and (b) activation scores.

SailAlign [31]. Next, we use the *TSYLB* toolkit [32] to break down each word into syllables. We derive the syllable time boundaries by creating a mapping between phoneme and syllables. Finally, we count the number of *syllables per second* (SPS) to get the reference syllable rate values. The process is implemented with scripts. While the reference scores are not free from errors, we manually inspected a subset of the sentences observing reasonable syllable boundaries. We quantify performance using the *mean squared error* (MSE) between the reference values and the estimates from the algorithms:

$$\text{MSE}\% = \frac{||\text{Reference rate} - \text{Estimated rate}||^2}{||\text{Reference rate}||^2} \cdot 100\% \quad (1)$$

Figures 1 shows the median MSE values achieved by the syllable rate estimation algorithms as function of valence and activation scores. We observe a drop in performance for turns perceived with lower valence and activation scores (e.g., sadness or boredom). For valence, Figure 1(a) shows that both algorithms achieved the best performance for positively-valenced sentences. For activation, Figure 1(b) shows that the best performance is achieved for sentences with medium level of arousal (neutral speech). For emotional speech, the algorithm proposed by Wang and Narayanan [19] gives better performance than the one proposed by de Jong and Wempe [20].

Table 1. Low level descriptors from speech. The derivatives of these LLDs are estimated and included for analysis [15].

	Low level descriptors	Nomenclature
Energy	Sum of RASTA style Auditory Spectrum	SumAudSpecRasta
	Sum of Auditory Spectrum	SumAudSpec
	RMS Energy	RMSenergy
	Zero Crossing Rate	ZCR
F0	Fundamental frequency	F0
	Probability of Voicing	ProbVoicing
V. Qua.	Jitter (Local)	JitterL
	Jitter (Delta)	JitterD
	Shimmer (Local)	ShimmerL
RASTA	Rasta-Style Filtered- Auditory Spectral bands[1-26]	VRasta[1-26]
Spectral	Spectral Flux	SpectFlux
	Spectral Entropy	SpectEn
	Spectral Variance	SpectVar
	Spectral Skewness	SpectSkew
	Spectral Kurtosis	SpectKurt
	Spectral Slope	SpectSlope
	Spectral Rolloff 0.25	SpectROff25
	Spectral Rolloff 0.50	SpectROff50
	Spectral Rolloff 0.75	SpectROff75
	Spectral Rolloff 0.90	SpectROff90
	Spectral Energy 25-650 Hz	Spectfband 25-650
	Spectral Energy 1k-4kHz	Spectfband 1k-4kHz
MFCC	Mel-frequency cepstrum coefficients	mfcc

4. EMOTION RECOGNITION EVALUATION

The second question that this study addresses is whether syllable rate features provide complementary, discriminative information over other acoustic features commonly used in current emotion recognition systems. We address this question by conducting controlled speech emotion classification experiments, aiming to quantify the improvement in performance achieved by adding syllable rate features.

4.1. Acoustic Features

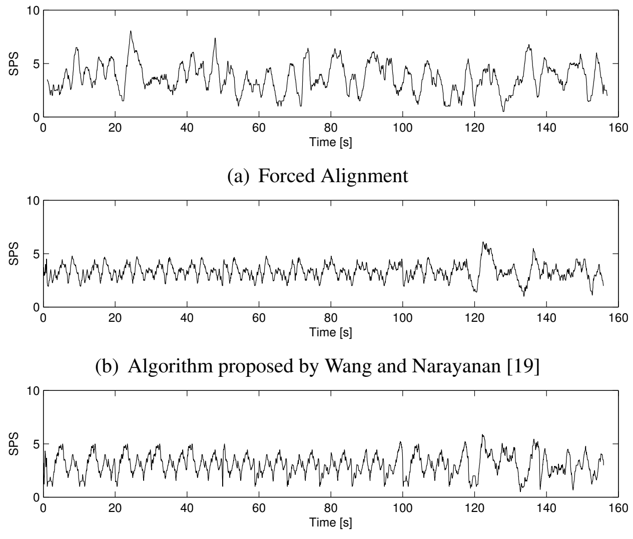
The externalization of emotion affects different speech characteristics. Therefore, current speech emotion recognition systems use large feature sets comprising various acoustic cues (e.g., spectral, prosodic and voice quality features). A popular approach consists in estimating *low level descriptors* (LLDs) such as fundamental frequency, and *Mel-frequency cepstral coefficients* (MFCCs). Next, statistics or functional such as mean and variance are estimated for each speech segment. These global statistics, referred to as *high level descriptors* (HLDs), are used as features of classifiers [33]. The present study relies on this framework. In particular, we use the feature set proposed for the Speaker State Challenge in

Table 2. High level descriptors derived from LLDs. The last four rows are statistics estimated only from the F0 [15].

Functionals	suffix
Quartiles 1-3	qrtl 1-3
Inter-quartile ranges	iqr1-2, iqr2-3, iqr1-3
Percentile (1%,99%)	prctl1.0, prctl99.0
Arithmetic Mean, Standard deviation	amean, std
Skewness, Kurtosis	skew, kurt
Mean of peak distances	meanPeakDist
Standard Deviation of peak distances	peakDistStd
Mean of peaks	peakMean
Arithmetic Mean of mean peaks	peakMMDist
Linear Regression Slope and Quadratic error	linregc1, linregerrQ
Quadratic Regression coef. and Quadratic error	qregc1, qregc2, qregerrQ
Contour Centroid	centroid
Duration when Signal below 25% range	dltime25
Duration when Signal above 90% range	ultime90
Duration when Signal risingfalling	risetime , falltime
Gain of linear prediction (LP)	lpgain
LP Coefficients	lpc 0-4
Percentage of non-zero frames	nnz
mean, max of segment length	meanSegLen, maxSegLen
min, std. dev. of segment length	minSegLen,StdsegLen
Input duration in seconds	duration

Interspeech 2011 [15]. The set includes 59 LLDs listed on Table 1. To understand better the contribution of speech rate features, we group these LLDs into six classes following the study of Busso and Rahman [34]: energy, F0 (fundamental frequency), voice quality, RASTA, spectral and MFCC (see Table 1). The first two groups correspond to suprasegmental prosodic features. The last three groups correspond to spectral features. Table 2 lists the HLDs estimated from these LLDs, which include 33 base functionals and 6 F0 functionals, forming a 4,368D sentence-level feature vector. The F0 functionals listed in the last four rows of Table 2 are the only duration features. Notice that they only describe the duration of voiced segments, not actual speech rate.

We follow a similar approach to derive features characterizing speech rate (i.e., defining sentence level statistics from LLDs). First, we create a speech rate profile for each speech segment (i.e., a LLD). We create this profile by estimating the syllable rate over 2s windows with 20 ms steps, producing 50 values per second (the algorithms to estimate syllable rate require at least 2s of speech to provide reliable estimations). We smooth the speech rate profiles using a median filter of order 10. This profile is estimated over the entire dialog, before segmenting it into speech turns. Figure 2 shows examples of syllable rate profiles estimated with forced alignment (i.e., reference values), and the two algorithms considered in this study. We estimate the following turn-level statistics (i.e., HLDs) from the profile: mean, variance, standard deviation, kurtosis, skewness, maximum, minimum, median, and upper and lower quartiles. This approach generates a 10D feature vector.



(a) Forced Alignment

(b) Algorithm proposed by Wang and Narayanan [19]

(c) Algorithm proposed by de Jong and Wempe [20]

Fig. 2. Speech rate profiles. The figures give the *syllable per second* (SPS) values across time (for one recording).

4.2. Experimental Settings

The main idea of this evaluation is to quantify the increase or drop in performance achieved by adding the proposed speech rate features to each of the six feature groups listed in Table 1: energy, F0 (fundamental frequency), voice quality, RASTA, spectral and MFCC. The tasks consist of binary classification problems for valence and activation. For each of these dimensions, we identify the median value assigned to all the speaking turns. Sentences with values higher than their median are assigned to one class, while the rest of the turns are assigned to the second class (i.e., low versus high valence; low versus high activation). This approach generates balanced classes, so the performance at chance level is 50%. For consistency, all the classifiers are trained with 50 features, as explained below.

Baseline classifiers: We build six baseline classifiers, one for each feature group. These classifiers are only trained with the HLDs derived from the LLDs belonging to the corresponding groups. We use a two-layer feature selection approach to reduce the feature dimension for the classifiers. In the first layer, we use *correlation feature selection* (CFS) to limit the number of HLDs per group to 400 features. CFS selects new features that are correlated with the class label, but that are not highly correlated with previously selected features. This approach is efficient and general, since it does not depend on the performance of any classifier. In the second layer, we use *forward feature selection* (FFS) to further reduce the feature set to 50 features by maximizing the performance of the classifiers.

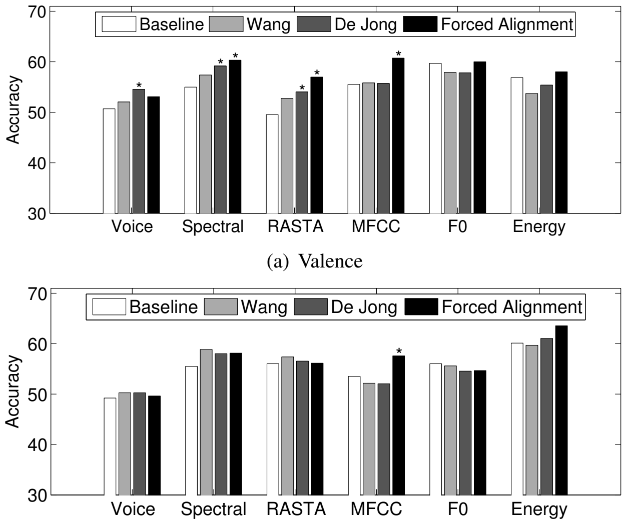
Classifiers with speech rate features: The baseline classifier for each group is compared with the ones trained by adding speech rate features to the set. We implement the following approach to limit the set to 50 features, per classifier. For a given group, we create an initial feature set with the 10 speech rate features described in Section 4.1. Then, we use FFS to increase the set to 50 features. The selected features complement the emotional information provided by speech rate features. For consistency, we only consider features that were previously selected by CFS (first layer of feature selection). This approach forces the classifiers to use speech rate features, but preserve the dimension of the feature set. Following this approach, we create three alternative classifiers per feature group by using the speech rate features estimated from forced alignment (i.e., reference), and the two syllable rate algorithms considered in this study.

All the classifiers are implemented with *support vector machine* (SVM) with linear kernel, using *sequential minimal optimization* (SMO). We rely on the implementation provided by *WEKA*. We use a three-fold cross-validation approach with balanced, speaker-independent partitions, where two groups are used for training and the remaining group for testing.

4.3. Classification Results

Figure 3 gives the accuracy of the binary classifiers trained for valence and activations. We added an asterisk on top of the bars to highlight cases where the improvement in accuracy with respect to their baseline classifiers is statistically significant. For this purpose, we use the large sample proportion hypothesis test, and we assert significance if p-value < 0.05.

When using syllable rate features, we observe higher improvements in accuracy for valence (Fig. 3(a)) than for activation (Fig. 3(b)). The groups with spectral features (i.e., *Spectral*, *RASTA* and *MFCC*) are the ones with higher improvements. Speech rate features complement the emotion information conveyed by spectral features. The accuracy does not significantly increase for groups conveying prosodic features (i.e., *F0* and *Energy*). The F0 group includes statistics about the duration of voiced segment which may be redundant with the syllable rate features. In general, the best performance is achieved when we use the reference syllable rate derived from forced alignment. This result demonstrates that it is important to develop syllable rate estimators that are robust against emotional speech.



(a) Valence

(b) Activation

Fig. 3. Accuracy of binary emotion classifiers for valence and activation by adding syllable rate features to different acoustic feature groups.

5. CONCLUSIONS

The study evaluated the performance of two syllable rate estimation algorithms in emotional speech. The analysis revealed a drop in accuracy for sentences perceived with low level of valence or activation. We carefully implemented emotion classification evaluations to explore whether the information provided by syllable rate features complements the information provided by other acoustic features. The results demonstrate that syllable rate features provide complementary information, especially for spectral features. However, the benefit of using speech rate features strongly depends on the quality of the estimations. This study suggests that advances on robust speech rate estimations will benefit speech emotion recognition systems.

The emotional content in the SEMAINE database includes subtle emotions. Given the recording settings, there are very few turns with strong emotional reactions. In our future work, we will replicate this analysis on other emotional databases with more extreme, or prototypical emotions. We will also consider other speech rate algorithms which may achieve better accuracy on emotional speech.

6. REFERENCES

[1] R. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[3] C. Busso, M. Bulut, and S.S. Narayanan, “Toward effective automatic recognition systems of emotion in speech,” in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds., pp. 110–127. Oxford University Press, New York, NY, USA, November 2013.

[4] K.R. Scherer, “Vocal communication of emotion: A review of research paradigms,” *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[5] I. O’Brien, N. Segalowitz, B. Freed, and J. Collentine, “Phonological memory predicts second language oral fluency gains in adults,” *Studies in Second Language Acquisition*, vol. 29, no. 4, pp. 557–581, December 2007.

[6] J.L. Miller, F. Grosjean, and C. Lomanto, "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," *Phonetica*, vol. 41, no. 4, pp. 215–225, 1984.

[7] W. Apple, L.A. Streeter, and R.M. Krauss, "Effects of pitch and speech rate on personal attributions," *Journal of Personality and Social Psychology*, vol. 37, no. 5, pp. 715–727, May 1979.

[8] B.L. Smith, B.L. Brown, W.J. Strong, and A.C. Rencher, "Effects of speech rate on personality perception," *Language and Speech*, vol. 18, no. 2, pp. 145–152, April 1975.

[9] M. Richardson, M. Hwang, A. Acero, and X. Huang, "Improvements on speech recognition for fast talkers," in *European Conference on Speech Communication and Technology (Eurospeech 1999)*, Budapest, Hungary, September 1999, pp. 411–414.

[10] J. Zheng, H. Franco, and A. Stolcke, "Rate-dependent acoustic modeling for large vocabulary conversational speech recognition," in *ISCA Tutorial and Research Workshop (ITRW) on Automatic Speech Recognition (ASR): Challenges for the New Millennium*, Paris, France, September 2000, ISCA, pp. 145–149.

[11] D.S. Pallett, J.G. Fiscus, W.M. Fisher, J.S. Garofolo, B.A. Lund, and M.A. Przybocki, "1993 benchmark tests for the ARPA spoken language program," in *Workshop on Human Language Technology (HLT 1994)*, Plainsboro, NJ, USA, March 1994, pp. 49–74.

[12] S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S.S. Narayanan, "An acoustic study of emotions expressed in speech," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 2193–2196.

[13] I.R. Murray and J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, February 1993.

[14] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of the Artificial Neural Networks in Engineering (ANNIE 1999)*, St. Louis, MO, November 1999.

[15] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, August 2011, pp. 3201–3204.

[16] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit - searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, January 2011.

[17] F. Dellaert and T. Polzin A. Waibel, "Recognizing emotion in speech," in *International Conference on Spoken Language (ICSLP 1996)*, Philadelphia, PA, USA, October 1996, vol. 3, pp. 1970–1973.

[18] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 1, pp. 1062–1087, December 2011.

[19] D. Wang and S.S. Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, November 2007.

[20] N.H. de Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, May 2009.

[21] C. Busso, S. Lee, and S.S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2225–2228.

[22] R. Faltlhauser, T. Pfau, and G. Ruske, "On-line speaking rate estimation using Gaussian mixture models," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*, Istanbul, Turkey, June 2000, vol. 3, pp. 1355–1358.

[23] Y. Zhang and J.R. Glass, "Speech rhythm guided syllable nuclei detection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, Taipei, Taiwan, April 2009, pp. 3797–3800.

[24] J. Yuan and M. Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 4222–4225.

[25] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in *European Conference on Speech Communication and Technology (Eurospeech 1997)*, Rhodes, Greece, September 1997, pp. 2079–2082.

[26] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998)*, Seattle, WA, USA, May 1998, vol. 2, pp. 729–732.

[27] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam, Amsterdam, Netherlands, 1996, http://www.praat.org.

[28] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[29] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, ISCA, pp. 19–24.

[30] J.R.J. Fontaine, K.R. Scherer, E.B. Roesch, and P.C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.

[31] A. Katsamanis, M. P. Black, P.G. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA, USA, January 2011.

[32] W.M. Fisher, "tsylb: NIST syllabification software, version 2 revision 1.1.," http://www.itl.nist.gov/iad/mig/tools/, 1997.

[33] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2253–2256.

[34] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.