EMOTION RECOGNITION USING SYNTHETIC SPEECH AS NEUTRAL REFERENCE

Reza Lotfian and Carlos Busso

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

ABSTRACT

A common approach to recognize emotion from speech is to estimate multiple acoustic features at sentence or turn level. These features are derived independent of the underlying lexical content. Studies have demonstrated that lexical dependent models improve emotion recognition accuracy. However, current practical approaches can only model small lexical units like phonemes, syllables or few key words, which limits these systems. We believe that building longer lexical models (i.e., sentence level model) is feasible by leveraging the advances in speech synthesis. Assuming that the transcript of the target speech is available, we synthesize speech conveying the same lexical information. The synthetic speech is used as a neutral reference model to contrast different acoustic features, unveiling local emotional changes. This paper introduces this novel framework and provides insights on how to compare the target and synthetic speech signals. Our evaluations demonstrate the benefits of synthetic speech as neutral reference to incorporate lexical dependencies in emotion recognition. The experimental results show that adding features derived from contrasting expressive speech with the proposed synthetic speech reference increases the accuracy in 2.1% and 2.8% (absolute) in classifying low versus high levels of arousal and valence, respectively.

Index Terms— emotion detection, synthetic speech, speech rate, speech alignment

1. INTRODUCTION

Emotions play an important role in human communications [1]. Detecting the underlying emotional state of the user can be valuable in the design of effective and engaging human computer interfaces (HCI). We externalize emotions by modulating multiple acoustic features. Prosodic features such as energy, fundamental frequency, and speech rate are affected by emotions [2–4]. Spectral features also present emotion dependent patterns. The externalization of emotion is a complex process that is tightly coupled with the underlying lexical content [5, 6]. For example, emotional changes on the first and second formants depend on the specific phoneme, where low vowels, such as /aa/, with less restricted tongue position present stronger emotional modulation than high vowels, such as /iy/ [7]. This dependency on the lexical content is commonly ignored in current emotion recognition systems [8]. However, studies have shown that lexical dependent models are more effective for emotion recognition [5, 9–13].

One of the main challenges in building lexical dependent models is defining the appropriate speech unit. Given the sparsity in emotional data, most of the studies have considered small speech units such as phonemes or syllables [5, 9, 10]. However, it is challenging to model suprasegmental variations characteristic of expressive speech with small speech units. Chauhan et al. [11] and Arias et al. [13] proposed sentence level models (i.e., one model per utterance). However, this approach is not practical in real applications. Is it possible to build practical, lexical-dependent, emotional models at sentence level? This paper proposes a novel solution based on synthetic speech and neutral reference models to address this problem.

Advances in speech synthesis provide an opportunity to create neutral reference models that can be contrasted with expressive speech. Since *text-to-speech* (TTS) systems are built with neutral speech, we expect that synthetic speech can provide a good representation of neutral speech. These reference models can be used to compare the acoustic properties of the target sentence. First, we synthesize speech with the same transcript of the target speech, and, therefore, they convey the same lexical information. We use word alignments and *dynamic time warping* (DTW) to temporally align both signals. We extract multiple features from the synthetic and target speech signals, which are directly compared producing relative features. The experimental evaluation demonstrates that adding these features increases the classification performance of the system in detecting sentences with low or high level of valence (negative versus positive) and arousal (calm versus active).

2. MOTIVATION

We have explored the powerful, scalable and appealing concept of using neutral reference models to contrast deviations in speech properties associated with emotions [13–16]. Detecting focal regions is required for success in designing truly novel robust emotion recognition systems. Our previous work considered lexical independent reference models for prosodic features in the forms of *hidden Markov models* (HMMs) [14], *Gaussian mixture models* (GMMs) [15] and *functional data analysis* (FDA) [16]. Extending the scope of these approaches, this paper proposes synthetic speech as a viable framework to create lexical dependent reference models. These models are used not only to contrast prosodic features, but spectral and voice quality features.

There are two underlying assumptions in this study. First, we assume that acoustic features derived from synthetic speech will provide a good representation of neutral speech. This assumption is reasonable, since TTS systems are trained with neutral speech, and their ultimate goal is to produce speech that is perceived as natural as possible. Second, we assume that lexical information (i.e., transcriptions) for the target sentences is available. While this assumption is valid in some practical scenarios, the use of *automatic speech recognition* (ASR) may be required. This exploratory study considers actual transcriptions to evaluate the potential of this approach.

This work was funded by NSF (IIS 1329659).

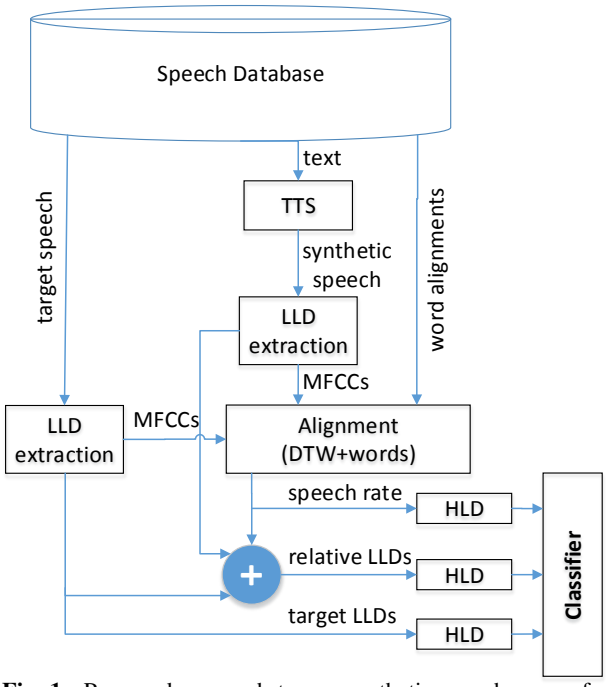


Fig. 1. Proposed approach to use synthetic speech as a reference model for emotion recognition.

While previous studies have used synthetic speech for training emotion recognition systems [17, 18], the use of synthetic speech as neutral reference to contrast expressive speech is novel.

3. APPROACH

This section describes the method proposed to use synthetic speech as a neutral reference to contrast the acoustic properties of the target speech. Details on the implementation are given in Section 4. Figure 1 shows the block diagram where *low level descriptors* (LLDs) correspond to acoustic features estimated frame-by-frame (e.g., F0 contour, RMS energy, *Mel-frequency cepstral coefficients* (MFCCs)). For each speaking turn, we have the audio, the transcription, and the word boundaries derived from forced alignment. A TTS system takes the transcription and generates a synthetic signal conveying the same lexical information as the target speech.

The synthetic and target speech signals are not timely aligned. Therefore, we use the word boundaries and DTW to estimate the best alignment. Section 3.1 describes this process. The final step is to contrast acoustic features of the target sentence with the corresponding ones extracted from synthetic speech. This process generates relative features that are added to the speech emotion classifier. Instead of using LLDs, we estimate statistics or functional from LLDs (e.g., the mean of the F0 contour), which are referred to as *high level descriptors* (HLDs). Sections 3.2 describes how we contrast target speech using the synthetic signal.

3.1. Time Alignment between Synthetic and Target Speech

The key idea of the approach is to compare frame-by-frame LLDs derived from the target and synthetic speech signals. Therefore, it is important to estimate the time alignment between both signals, which is implemented in two steps. The first step takes the word boundaries of both signals either derived from forced alignment (target speech) or provided by the TTS system (synthetic speech). The second step uses DTW to estimate alignment within words boundaries (i.e., sub-word alignment). DTW uses MFCCs as features for the alignment, which have been successfully used in DTW-based frameworks for discrete word recognition [19], and connected word recognition [20]. The allowable region for the dynamic path was set within the range of [1/3,3].

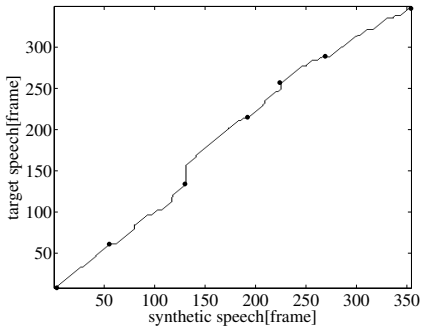


Fig. 2. Example of alignment path between synthetic speech and target speech for one sentence. The dots correspond to the word boundaries.

Figure 2 shows the warping path over one utterance. The dots within the path correspond to constraints imposed by word boundaries in both signals (first step). Therefore, the alignment path always includes the word boundaries of target and synthetic speech signals. This alignment path is used to estimate new relative features for speech emotion recognition, as described next.

3.2. Contrasting Target and Synthetic Speech Signals

Using the alignment path, we estimate two sets of features that capture the differences between our reference neutral model (synthetic speech) and the target speech. The first set corresponds to speech rate features. The TTS system generates synthetic speech using duration models that are appropriate for neutral speech. Emotional speech has characteristic speech duration patterns that deviate from neutral speech [4]. Here, we aim to use synthetic speech to identify emotional cues reflected in longer or shorted speech durations from the one that we would expect from neutral speech. We generated a LLD by estimating localized ratio between the number of frames from one signal corresponding to frames of the other signal. Then, we convert the speech rate signal into logarithmic scale. Finally, we smooth the resulting curve using a low pass Hamming filter of length 0.5 seconds. Figure 3 gives an example of the speech rate curve generated with this approach, which we consider as an extra LLD. The HLDs derived from this curve are referred to as *Duration HLD*.

The second set of features corresponds to relative frame-by-frame features estimated as follows. We compute LLDs for the target and synthetic speech signals. Then, we align these feature vectors by using the utterance level warping path. When one frame of one signal is assigned to multiple frames of the other, we estimate the average value of individual LLDs across frames. After the LLDs are aligned, we subtract their values. This set of features reflects the deviations, in the feature space, of the target speech from the ones derived from our reference neutral model – the synthetic speech. We expect that these differences will increase as the acoustic features of the target signal deviate from the expected patterns observed in neutral speech. Since we estimate these relative features after time-alignment, the lexical dependency is intrinsically captured

by this approach. These relative features can be used as extra LLDs for speech emotion recognition. Similar to other features, we can apply HLDs over speech segments, capturing statistics of these relative features. The HLDs derived from these features are referred to as *relative HLD*.

4. IMPLEMENTATION

4.1. Database

The study relies on the SEMAINE database, which includes natural human interactions between a *user* and an *operator* [21]. Using the *sensitive artificial listener* (SAL) framework, the operator plays different characters with specific personalities, inducing emotional reactions on the user. While the corpus provides audiovisual recordings, this study only considers the audio from ten users.

The duration of the sessions is approximately five minutes and are evaluated by six raters in terms of arousal (calm versus active), valence (negative versus positive), power (weak versus strong), and expectation (predictable versus unexpected). We only consider arousal and valence. The emotional labels are annotated using FEELTRACE [22], which provides time-continuous traces reporting frame-by-frame the emotional perception of the evaluators as they watch the interactions. This study considers two separate binary classification problems consisting in detecting sentences with low or high level of arousal and valence. First, we correct the time-continuous emotional labels by modeling the reaction lag of the evaluators following the approach proposed by Mariooryad and Busso [23, 24]. Then, we segment the corpus into fix windows of 1 sec, without overlap. For each emotional dimension, we estimate the average value of the traces across evaluators during the duration of each segment. Finally, we use the median of the values to define the binary classes (low versus high) for arousal and valence. In total, we consider 10,799 1-sec segments.

4.2. Speech Synthesis

As shown in figure 1, the system takes an input speech with its transcription. The transcription is used to synthesize speech conveying the same lexical content as the target sentence. The TTS system used in this study is Festival, which is a general multi-lingual speech synthesis system [25]. In particular, we use cluster unit selection trained for an American male speaker as our speech synthesis approach [26], which is based on the concatenation of sub-word units from a database of labeled speech. The appropriate units are selected based on the phonetic and prosodic content by finding the optimal path through the candidate units.

we use cluster unit selection as our speech synthesis approach

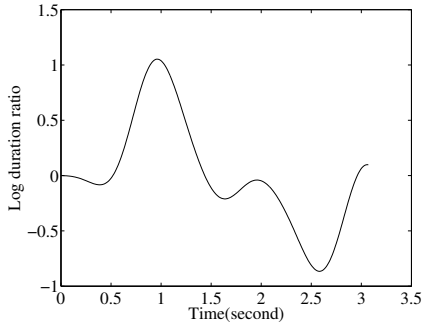


Fig. 3. Smoothed speech rate curve for one utterance. The curve gives the localized ratio between the frame durations of the synthetic and target speech signals, expressed in logarithmic scale.

4.3. Acoustic Feature

We implement the approach using the LLDs proposed for the Speaker State Challenge at Interspeech 2011 [27]. The feature set is estimated with openSMILE [28], and consists of 60 different acoustic features extracted every 10ms. The set includes MFCCs, RASTA features, F0 contour, RMS energy, and voice quality features. Instead of using the HLDs in the proposed set (4368 features derived from this LLDs), we simplify the approach by estimating only 17 HLDs: arithmetic mean, minimum, maximum, standard deviation, kurtosis, skewness, number of zero crossing, number of mean crossing, number of max, number of min, mean value of max peaks, mean value of min peaks, linear regression slope, linear regression error, quadratic regression coefficients (a and b) and quadratic error. The experiments considers three features sets:

- *Baseline HLD*: Target sentences – functionals derived from LLDs and their first order derivative (2040=60×2×17)
- *Duration HLD*: Target and synthetic speech signals (Sec. 3.2) – functionals derived from this duration LLD and its first order derivative (34=1×2×17)
- *Relative HLD*: Target and synthetic speech signals (Sec. 3.2) – functionals derived from relative LLDs and their first order derivative (2040=60×2×17)

Given the high dimension of the features, we consider a two-layer feature selection approach for each of the conditions evaluated in this study. The goal of the first layer is to efficiently reduce the set. We use information gain ratio for this purpose, which reduces the number of feature to 500. The second layer is implemented with *correlation feature selection* (CFS). CFS adds features one-by-one by maximizing the correlation between the features and the labels, and by minimizing the correlation between selected features. We reduce the feature dimension to 100 for all the experiments, with the exception of the evaluation of the *Duration HLD* set, which has only 34 features.

4.4. SVM Classifier

The emotion discrimination of the proposed relative features is evaluated with *support vector machine* (SVM) trained with *sequential minimal optimization* (SMO) (implemented with WEKA [29]). We use a polynomial kernel, where the SVM complexity parameter is set to c=0.1. The evaluation considers the *leave-one-subject-out* (LOSO) cross-validation approach, where all the data from one subject is either in the training or in the testing set (speaker-independent partitions). Since we consider balanced emotional classes (Sec. 4.1), we report the weighted average accuracy across speaker (ten folds).

5. RESULTS

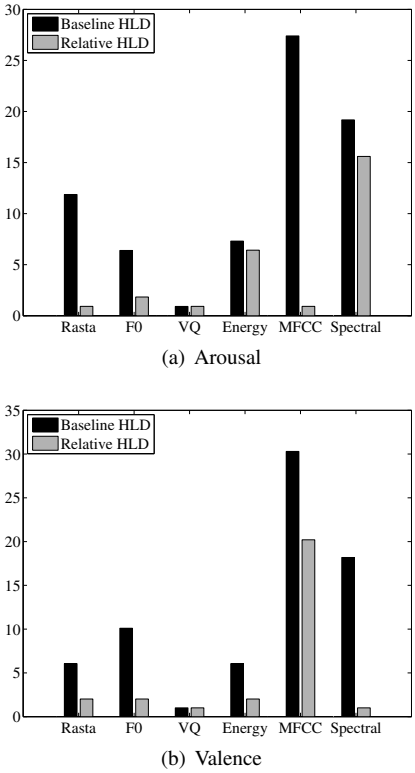
We evaluate the SVM classifier with different set of features. Table 1 lists the results. A classifier trained with the *Baseline HLDs* achieves 63.8% and 59.9% of accuracy in detecting low and high levels of arousal and valence, respectively. Notice that this is a challenging database, where it is very difficult to improve classification performance. When we use the *Duration HLDs* set, the classification performance is better than chances. Notice that this set only has 34 features, so we do not implement feature selection. It is interesting

Table 1. Classification accuracy for low versus high level of arousal (*Aro.*) and valence (*Val*). The classification performance at random is %50, since the binary classes are balanced.

Feature set	Aro. [%]	Val. [%]
Baseline HLD	63.8	59.9
Duration HLD	57.4	56.8
Relative HLD	63.3	58.6
Relative HLD & Duration HLD	63.5	60.3
Baseline HLD & Relative HLD	64.1	61.6
Baseline HLD & Relative HLD & Duration HLD	66.0	62.7

that contrasting the speech rate of both signals is effective in recognizing emotions. The duration models in the TTS systems offer an opportunity to obtain discriminative emotional features with the proposed framework. When we train the classifiers with the *Relative HLD* set, the performance of the system is similar to the classifier trained with the *Baseline HLDs*. The best performance is achieved when all the feature sets are simultaneously considered (last row in Table 1). The classifier trained with these features outperforms the classifier only trained with the *Baseline HLDs* by 2.1% and 2.8% (absolute) for arousal and valence, respectively. The relative features derived after contrasting the acoustic features of the target and synthetic speech signals provide complementary information that can improve the classification performance of the speech emotion recognition system.

We evaluate the features selected for the classifier trained with *Baseline HLDs* and *Relative HLD*. For the analysis, we group the LLDs into six classes, following the study of Busso and Rahman [30]: RASTA, F0 (fundamental frequency), voice quality (VQ), energy, MFCC and spectral features. While RASTA and MFCCs features are spectral features, their HLDs are grouped into separate classes (i.e., they are not included in the class "spectral"). As described in Section 4.3, we consider only 100 features from both sets after the feature selection step. Figure 4 shows the number of features included per group, for *Baseline HLDs* and *Relative HLDs*. Over 26% of the selected features correspond to *Relative HLD*. This result clearly demonstrates the contribution of the proposed features in discriminating expressive speech. Most of the selected features for arousal come from spectral features (very few MFCCs are selected). We observe the opposite pattern for valence. Most of the features come from MFCCs, and very few from spectral features. We are currently exploring the underlying reasons for this result.



(a) Arousal

(b) Valence

Fig. 4. Features used to train the classifiers after feature selection over the *Baseline HLD* and *Relative HLD* sets. The figure shows the number of features selected from each feature set. The results are split into six feature groups.

6. CONCLUSION

This study introduced a novel approach to create lexical dependent models at utterance-level using synthetic speech. For a given target speech, we generate synthetic speech conveying the same lexical information. We use this signal as a neutral reference model to contrast expressive speech, deriving relative features that capture deviations from neutral speech. The experimental evaluation demonstrates that these features provide complementary information improving the performance of the system in 2.1% and 2.8% (absolute) in classifying low versus high levels of arousal and valence, respectively.

This exploratory study on synthetic speech for emotion recognition opens several research directions. The framework to contrast the target and synthetic speech can be improved. For example, instead of warping the features, we can re-synthesize the sentences using the correct time boundaries. Therefore, the synthetic signal not only will convey the same information as the target sentences, but also will be timely synchronized, facilitating direct comparison. Likewise, we are evaluating different speech synthesis approaches to create the most suitable neutral reference model. Finally, instead of building a single synthetic speech, we are planning to build a family of synthetic sentences using different TTS methods. We expect that this family of sentences will provide a better characterization of neutral speech to detect deviations caused by localized emotional behaviors.

Acknowledgements

Portions of the research in this paper use the Semaine Database collected for the Semaine project (www.semaine-db.eu) [21]

7. REFERENCES

[1] 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.

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

[3] C. Busso, M. Bulut, S. Lee, and S.S. Narayanan, "Fundamental frequency analysis for speech emotion processing," in *The Role of Prosody in Affective Speech*, Sylvie Hancil, Ed., pp. 309–337. Peter Lang Publishing Group, Berlin, Germany, July 2009.

[4] M. Abdelwahab and C. Busso, "Evaluation of syllable rate estimation in expressive speech and its contribution to emotion recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, South Lake Tahoe, CA, USA, December 2014, pp. 472–477.

[5] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, February 2014.

[6] J.H.L. Hansen and B.D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 4, pp. 307–313, July 1996.

[7] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S.S. Narayanan, "Emotion recognition based on phoneme classes," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 889–892.

[8] 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.

[9] B. Vlasenko, D. Prylipko, D. Philippou-Hübner, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011, pp. 1577–1580.

[10] B. Vlasenko, D. Philippou-Hübner, D. Prylipko, R. Böck, I. Siegert, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal emotions," in *IEEE International Conference on Multimedia and Expo (ICME 2011)*, Barcelona, Spain, July 2011.

[11] R. Chauhan, J. Yadav, S.G. Koolagudi, and K.S. Rao, "Text independent emotion recognition using spectral features," in *Contemporary Computing*, S. Aluru, S. Bandyopadhyay, U. Catalyurek, D. Dubhashi, P. Jones, M. Parashar, and B. Schmidt, Eds., vol. 168 of *Communications in Computer and Information Science*, pp. 359–370. Springer-Verlag Berlin Heidelberg, Berlin Heidelberg, January 2011.

[12] L. Fu, X. Mao, and L. Chen, "Relative speech emotion recognition based artificial neural network," in *Pacific-Asia Workshop on Computational Intelligence and Industrial Application (PACIIA 2008)*, Wuhan, China, December 2008, vol. 2, pp. 140–144.

[13] J.P. Arias, C. Busso, and N.B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech and Language*, vol. 28, no. 1, pp. 278–294, January 2014.

[14] 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.

[15] C. Busso, S. Lee, and S.S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[16] J.P. Arias, C. Busso, and N.B. Yoma, "Energy and F0 contour modeling with functional data analysis for emotional speech detection," in *Interspeech 2013*, Lyon, France, August 2013, pp. 2871–2875.

[17] B. Schuller, Z. Zhang, F. Weninger, and F. Burkhardt, "Synthesized speech for model training in cross-corpus recognition of human emotion," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 313–323, September 2012.

[18] B. Schuller and F.Burkhardt, "Learning with synthesized speech for automatic emotion recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 5150–5153.

[19] L.R. Rabiner, A.E. Rosenberg, and S.E. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 6, pp. 575–582, December 1978.

[20] C.S. Myers and L.R. Rabiner, "A comparative study of several dynamic time-warping algorithms for connected-word recognition," *Bell System Technical Journal*, vol. 60, no. 7, pp. 1389–1409, September 1981.

[21] 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.

[22] 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.

[23] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. To appear, 2014, Special Issue Best of ACII.

[24] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 85–90.

[25] P. Taylor, A. W. Black, and R. Caley, "The architecture of the festival speech synthesis," in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, Blue Mountains, NSW, Australia, November 1998, pp. 147–151.

[26] A.W. Black and P.A. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *EUROSPEECH 1997*, Rhodes, Greece, September 1997, pp. 601–604.

[27] 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.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, June 2009.

[30] 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.