Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech

Carlos Busso, *Member, IEEE,* Soroosh Mariooryad, *Student Member, IEEE,*
Angeliki Metallinou, *Student Member, IEEE,* Shrikanth Narayanan, *Fellow, IEEE*

Abstract—The externalization of emotion is intrinsically speaker-dependent. A robust emotion recognition system should be able to compensate for these differences across speakers. A natural approach is to normalize the features before training the classifiers. However, the normalization scheme should not affect the acoustic differences between emotional classes. This study presents the *iterative feature normalization* (IFN) framework, which is an unsupervised front-end, especially designed for emotion detection. The IFN approach aims to reduce the acoustic differences, between the neutral speech across speakers, while preserving the inter-emotional variability in expressive speech. This goal is achieved by iteratively detecting neutral speech for each speaker, and using this subset to estimate the feature normalization parameters. Then, an affine transformation is applied to both neutral and emotional speech. This process is repeated till the results from the emotion detection system are consistent between consecutive iterations. The IFN approach is exhaustively evaluated using the IEMOCAP database and a dataset obtained under free uncontrolled recording conditions with different evaluation configurations. The results show that the systems trained with the IFN approach achieve better performance than systems trained either without normalization or with global normalization.

Index Terms—Emotion recognition, speaker normalization, emotion, features normalization.

1 Introduction

AUTOMATIC emotion recognition has the potential to change the way humans interact and communicate with machines. Some of the domains that could be enhanced by adding emotional capabilities include call centers, games, tutoring systems, ambient intelligent environments and health care. The promising results shown by the early technological developments in automatic emotion recognition however have not materialized into significant advances in real life applications. One of the main barriers toward detecting the emotional state of a human is the inherent inter-speaker variability found in emotional manifestations. It is in this context, we propose a novel normalization scheme designed to reduce the speaker variability, while preserving the discrimination between emotional states.

Data normalization is an important aspect that needs to be considered for a robust automatic emotion recognition system [1]. The normalization step should reduce all sources of variability, while preserving the differences between normal and emotionally expressive speech. In particular, it should compensate for speaker variability. Speech production is the result of controlled movements of an individual's vocal tract apparatus that include the lungs, trachea, larynx, pharyngeal cavity, oral cavity, and nasal cavity. As a result, the properties of speech are intrinsically speaker-dependent. For example, the fundamental frequency is physically constrained by the anatomy of the larynx, which explains some of the gender differences observed in speech. Likewise, a robust emotion recognition system should cope with differences in the recording conditions. The quality and property of the speech highly depend on the sensor and technology used to capture and transmit the speech (e.g., mobile versus landline systems, close-talking versus far-field environment microphones). Any mismatch in the recording condition between the training and testing speech sets will affect the features extracted from the signals. For instance, it is well-known that the energy tends to increase with angry or happy speech [2]. If the energy of the speech signal is not properly normalized, any difference in the microphone gain will affect the performance of the system (e.g., high signal amplitude speech may be confused with emotional speech).

Most of the current approaches to normalize speech or speech features are based on gross manipulation of the speech at utterance or dialog turn level. In many cases, either speech is not normalized or the approach is not clearly defined. Given the importance of the normalization step, the limited progress in this area is surprising. Some of the approaches that have been widely used are z-standardization (subtract the mean and divide by the standard deviation) [3], min-

C. Busso and S. Mariooryad are with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, TX 75080 (e-mail: busso@utdallas.edu, sxm096221@utdallas.edu).
A. Metallinou is with Pearson Knowledge Technologies, Menlo Park, CA 94025, USA (angeliki.metallinou@pearson.com)
S. Narayanan is with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA (shri@sipi.usc.edu)
Manuscript received November xx, xxxx; revised xxxx xx, xxxx.

max normalization (scaling features between -1 and 1) [4], and subtraction of mean values [1]. All these manipulations applied across speakers and speech samples affect the discrimination between emotional classes. An interesting approach was proposed by Batliner et al. [5]. For a given lexical unit (i.e., word or phoneme), the speech features were normalized by estimating reference values for "average speakers" learned from a training database. These reference values were used to scale the duration and energy of the speech. However, this normalization scheme does not cope well with inter-speaker variation.

In previous work, we had proposed a speaker-dependent normalization scheme [6], [7]. The main idea of the approach is to estimate the linear scaling parameters for each speaker based only on his/her neutral speech. Then, the normalization parameters are applied to all speech samples from that speaker, including the emotional speech set. Given that the normalization is an affine transformation, the scaling factors will not affect emotional discrimination in the speech, since the differences observed in the features across emotional categories will be preserved. The rationale behind this normalization is that, while individuals may express emotions differently, the patterns observed in their neutral utterances should be similar across speakers. The assumptions made in this normalization approach are that the identity of the subjects is known, and that the labels for a neutral speech set are available for each speaker. These assumptions are not realistic in many practical applications. This paper describes an *iterative feature normalization* (IFN) scheme to overcome this issue in the context of emotion detection [8], [9]. This unsupervised feature normalization approach extends the aforementioned ideas by iteratively detecting a neutral speech subset for each speaker, which is used to estimate his/her normalization parameters. The speaker-dependent scaling parameters are then applied to the entire corpus including expressive speech. The approach is implemented using z-normalization of statistics derived from acoustic features. Notice that the mean and standard deviations are only estimated from the detected neutral subset.

Although the IFN approach addresses the assumption that neutral speech is available to estimate the normalization parameters, it still requires the speaker identity (i.e., speaker-dependent normalization scheme). However, our results suggest that the normalization is not sensitive to errors in the speaker assignments (automatically determined). The experimental results demonstrate the potential of the proposed approach.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 introduces the proposed normalization scheme for emotion recognition. Section 4 describes the database and acoustic features used in this study. Section 5 presents the experimental design and results of the proposed normalization scheme. The paper concludes with Section 6, which gives the discussion, future work and final remarks.

2 Background

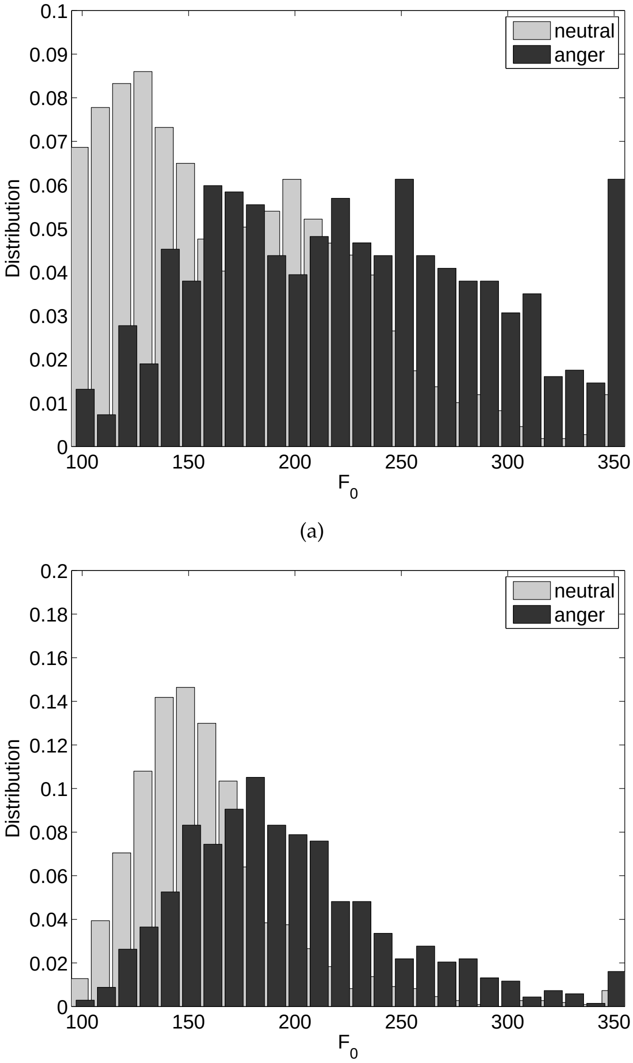
2.1 Motivation

A variety of acoustic features have been used to recognize emotions from speech [10], [11]. Features that are commonly used include the fundamental frequency, energy, formants, speech rate and voice quality features. These features are not only affected by the externalization of emotion, but also by the speaker and phonetic variabilities. Therefore, a normalization scheme is important to reduce variability while preserving the emotion-related patterns. This normalization step is critical for building a robust speech emotion recognition system.

Let us consider, for example, the fundamental frequency (F0) mean, which is widely used as a feature for emotion recognition. In fact, our previous analysis has indicated that the F0 mean is one of the most emotionally prominent aspects of the F0 contour, when properly normalized [6]. The fundamental frequency is directly constrained by the structure and size of the larynx and vocal folds [12]. The F0 contour for men is in the range 50-250Hz, while for women is higher (120-500Hz) [12]. Figure 1(a) shows the distribution of the F0 mean for neutral and angry sentences recorded from five men and five women (IEMOCAP corpus – Sec. 4.1). Although angry speech has higher F0 values than neutral speech [13], speaker variability increases the confusion between both emotional classes. As a result, emotional differences are blurred by inter-speaker differences. Figure 1(b) shows the same distribution after speaker normalization (see Sec 3). The overlap between both classes is now reduced. The shift in the distributions can be directly associated to emotional variations.

2.2 Related Work

Despite the importance of the feature normalization, as illustrated in Section 2.1, few studies have addressed this problem for emotion detection and recognition. Various types of normalization schemes can be implemented according to the target problem. For example, speaker normalization can be used to compensate for speaker-dependent variability. However, this scheme assumes that we know or can infer the speaker identity. Similarly, a normalization scheme can be used to compensate for phonetic (lexical) variability. This approach assumes that either the transcriptions or an *automatic speech recognition* (ASR) system is available. For multi-corpora studies, feature normalization can be designed to compensate for variability in recording conditions. This approach



(a)

(b)

Fig. 1. Speaker versus emotion variability. (a) F0 mean distribution without normalization, (b) F0 mean distribution with normalization (Sec. 3).

can prove to be useful for the deployment of emotion recognition system in real applications. This section discusses some of the most common approaches for feature normalization. Table 1 summarizes the feature normalization approaches proposed in previous studies.

A simple approach that requires no speaker or phonetic information is to normalize all the available features regardless of the speaker or lexical content. A common approach is z-standardization, in which the features are separately normalized by subtracting their mean and dividing by their standard deviation (i.e., zero mean, unit variance). These parameters are estimated across the entire data [3], [14], [15], [16], [17]. An alternative normalization method is min-max normalization, where features are scaled within a predefined range (e.g., from -1 to 1) [4], [16], [18]. Pao et al. [19] used zero mean normalization for all features. Another global feature normalization is the approach described by Yan et al. [20]. They proposed an exponential transformation to normalize speech features so that they follow a normal distribution. They showed that the approach generates features that improve the performance of emotion classification using *quadratic discrimination functions* (QDFs).

Previous studies have considered speaker-dependent normalizations. A common approach is to use z-standardization by computing speaker-specific means and variances [16], [21], [22], [23], [24]. The min-max normalization scheme can also be implemented by considering speaker-specific transformations [16], [25]. Wollmer et al. [16] compared these normalization approaches (z-standardization, min-max and combination of both techniques), implemented in a speaker-dependent and speaker-independent manner. Their results while instructive are not conclusive since the best normalization performance varies across the adopted machine learning approaches. Other speaker-specific normalization approaches include division of speech energy by its mean [1], [26], and the common cepstral mean subtraction of *Mel frequency cepstral coefficients* (MFCCs) [21], [27]. Sethu et al. [28] have described another interesting approach. They proposed a feature warping scheme that maps the initial feature distribution into a pre-determined distribution (i.e., standard normal distribution). They separately warp the features of each speaker, including his/her neutral and emotional samples. Their results indicate that the proposed speaker-dependent feature warping approach improves emotion recognition performance.

Schuller et al. [24] explored the case when the training and testing partitions were created from different corpora. For this problem, they normalized the features by estimating corpus-specific and speaker-specific z-standardization parameters. A speaker and lexical dependent normalization approach was presented by Fu et al. [31]. The scheme uses relative features that capture the differences of a feature with respect to a neutral reference. This reference is computed using neutral, speaker-specific and utterance-specific samples from the database. However, it is not clear how this approach can be generalized for cases where the information about speaker, emotion and lexical content is not available. Mariooryad and Busso [29] proposed a shape based lexical normalization scheme that uses the whitening transformation to compensate for the variability observed across different phonemes. Finally, Zeng et al. have presented an approach closest to this work [30]. Each feature is divided by the feature mean, computed per speaker from held-out samples of neutral speech. These samples are assumed to be available in advance.

We have shown in our previous work that global normalization is not always efficient in improving the

TABLE 1
Summary of feature normalization approaches for speech emotion recognition.

Method	References	Description
Global Normalization		
z-standardization	[3], [14], [15], [16], [17]	mean and variance of each feature is computed using all available data
min-max normalization	[4], [16], [18]	scaling of all feature values within a predefined range
min-max and z-standardization	[16]	combination of the above two schemes for all available data
zero mean normalization	[19]	mean value computed for all available data
exponential transformation	[20]	use a modified exponential transform so that the transformed features follow a normal distribution
Speaker-dependent normalization		
z-standardization	[16], [21], [22], [23], [24]	mean and variance of each feature is computed using speaker-specific data
min-max normalization	[16], [25]	scaling of speaker-specific feature values within a predefined range
min-max and z-standardization	[16]	combination of the above two schemes for speaker-specific data
divide energy by energy mean	[1], [26]	mean value computed per speaker
divide energy by energy peak	[27]	peak value computed per speaker's utterance
cepstral mean subtraction	[21], [27]	performed for spectral features such as MFCCs
feature warping	[28]	speaker-specific feature warping so that features follow a normal distribution
whitening normalization	[29]	speaker-specific feature normalization with whitening transformation of a shape-based representation
Speaker and corpus-dependent normalization		
z-standardization	[24]	mean and variance of features are computed using speaker-specific data when many corpora are available
Speaker and emotion-dependent normalization		
divide each feature by its mean	[6], [30]	mean is computed from speaker-specific neutral set which is assumed available in advance
Speaker, text and emotion-dependent normalization		
relative features	[31]	use of relative features which measure change with respect to a reference computed from neutral, speaker-specific and text-specific data

performance of an emotion recognition system [8]. We have argued that this type of normalization affects the discrimination between emotional classes. While estimating normalization parameters that are dependent on the underlying lexicon, speaker or emotional classes may give better performance, an unsupervised approach is needed for practical applications. This work proposes a robust unsupervised front-end to normalize the acoustic features for emotion detection. Although the approach is speaker-dependent, the evaluation results reveal that the scheme is not sensitive to errors made on the assumed speaker identity. This novel normalization approach produces higher accuracies in data obtained from both controlled and uncontrolled recording settings, as demonstrated in section 5.

3 Methodology

3.1 Ideal Feature Normalization

We have previously proposed a speaker-dependent normalization scheme to compensate for inter-speaker variability and recording conditions [6], [7]. The main idea is to reduce the differences observed in neutral speech across speakers, while preserving the emotional discrimination observed between emotional categories. This goal is achieved by separately estimating the normalization parameters for each speaker using only his/her neutral speech. For each subject, the estimated normalization parameters are then applied to his/her entire speech data, including the emotional set. These normalizations are performed such that the properties of their neutral speech become similar across speakers (e.g., first and second order statistics). If the normalization uses an affine transformation, the relative differences between neutral and emotional speech in the feature space will be preserved.

The motivation on this ideal normalization scheme is given in Figure 2. Figure 2(a) shows a schematic characterization of emotional clusters in the feature space for two subjects before normalization. Emotional classes are overlapped given the intrinsic speaker dependency. For example, neutral samples from speaker 1 are mixed with sad samples from speaker 2. Figure 2(b) describes the approach which aims to normalize the corpus such that neutral sets across speakers have similar properties. Figure 2(c) gives the clustering of the emotional classes in the feature space after this ideal normalization. Notice that this normalization approach is general and can be applied using different transformations (e.g., z-standardization, min-max normalization and feature warping). Likewise, it can be applied to other modalities such as facial features. For example, Zeng et al. [30] proposed to use neutral facial poses to normalize their facial features.

One assumption made in this approach is that neutral speech will be available for each speaker to estimate his/her normalization parameters. There are two major implications/limitations of this assumption: (i) the identities of the speakers are known, and (ii) reference neutral speech is available for each speaker. For real-life applications, these assumptions are reasonable when the speakers are known, and a few seconds of their neutral speech can be pre-recorded. For example, this speaker-dependent normalization scheme can be used for personalized interfaces with emotional capabilities designed for mobile devices.

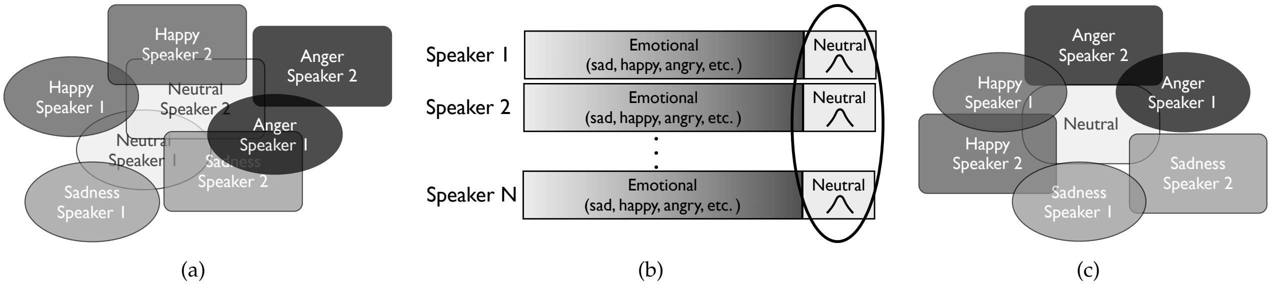


Fig. 2. Schematic description of ideal feature normalization from neutral speech. (a) emotional classes in feature space before normalization, (b) corpus is normalized such that neutral portions of the corpus have similar statistics across speakers, and (c) emotional classes in feature space after normalization.

However, in many applications either the identity of the speaker is unknown or neutral speech data are not readily available. In those cases, this normalization scheme cannot be implemented.

3.2 Iterative Feature Normalization (IFN)

Given the limitations of the ideal feature normalization approach, the present paper proposes the *iterative feature normalization* (IFN) framework. This scheme is an unsupervised front-end that overcomes the requirement of having prerecorded neutral sets for the test speakers. However, we assume that for the training speakers, neutral speech samples are provided, which is a reasonable assumption given that classifiers are built with emotional labels. For each training speaker the acoustic features are normalized with respect to the statistics extracted from their neutral samples (i.e., ideal normalization). Then, inspired by co-training strategies [32], the IFN approach iteratively detects the neutral speech set to estimate the normalization parameters for the test speaker. The parameters are then applied to all the samples of that speaker, including his/her emotional sentences, preserving the differences between emotional classes.

Figure 3 describes the IFN approach. The following procedure is separately implemented for each speaker in the test data. First, the normalization parameters are initialized. Then, the speech features are normalized using an affine transformation, which can be implemented using various approaches (e.g. z-standardization, see Sec. 3.3). Then, an emotional speech detection algorithm is used to discriminate between neutral and emotional speech (details are given in Section 3.3). The emotional labels assigned to the speech files are compared with the emotional labels from the previous iteration. If the percentage of emotional labels that are modified during the iteration is higher than a given threshold, a new iteration is computed. An alternative stopping criterion is setting a predefined maximum number of iterations. For each iteration, only the speech samples labeled as neutral are used to estimate the normalization parameters. The performance of the emotion detection system is expected to increase as the error estimation on the normalization parameters decreases. A better classification result will produce a better estimation of the neutral subset to estimate the normalization parameters.

Notice that the IFN scheme is a speaker-dependent normalization (i.e., the identity of the speaker is required). In some applications, it is safe to assume that only one speaker uses the system at a time (e.g., call center application). In other cases, we may be interested in tracking emotion in multi-person interaction (e.g., in smart room environment [33]). In these applications, the identity of the speakers will need to be predicted using either supervised or unsupervised speaker identification. However, the experiments in Section 5.3 reveal that the IFN approach is not very sensitive to errors made in identifying the speakers.

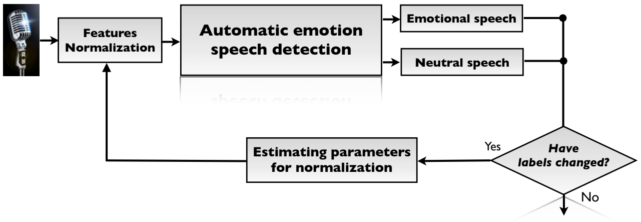


Fig. 3. Iterative feature normalization. This unsupervised front-end uses an automatic emotional speech detector to identify neutral samples, which are used to estimate the normalization parameters. The process is iteratively repeated until the labels are not modified any further (see Sec. 3.2 and 3.3).

3.3 Implementation

The proposed approach in Figure 3 is general and can be implemented with different affine transformations and emotion detection systems. In this paper, we implemented the approach with z-standardization and *support vector machine* (SVM).

Equation 1 describes the z-standardization approach. This affine transformation aims to preserve the first (mean) and second (variance) order statistics between the sentence level features derived from the neutral subsets across speakers. For a given speech feature from the speaker s, f^s, its mean value, $\mu_{neu}^{f^s}$, and standard deviation, $\sigma_{neu}^{f^s}$, are estimated using only the neutral samples. Then, the normalized feature $\widehat{f^s}$ is estimated as described in Equation 1.

$$\widehat{f^s} = \frac{f^s - \mu_{neu}^{f^s}}{\sigma_{neu}^{f^s}} \tag{1}$$

One important component of the proposed approach is the automatic emotional speech detector (Fig. 3). Unlike other emotion classification systems that recognize multiple emotional labels, the goal of this detection system is to identify neutral speech with high precision rate. This study uses a linear kernel SVM with *sequential minimal optimization* (SMO). Among many hyperplanes, SVM selects the one that has the largest margin between two classes (i.e., maximum margin classifier). We have successfully used this machine learning framework in paralinguistic recognition problems such as emotion recognition [9], [34] and sleepiness detection [35]. The SVM is trained and tested with the WEKA data mining toolkit [36]. For consistency across the evaluations, we set the complexity parameter of the classifier to 0.1. This value provided good performance in preliminary experiments.

An important aspect of the IFN approach is setting the values of the normalization parameters in the first iteration (i.e., initialization). These parameters can be initialized with different approaches. Given that it is not clear which approach provides better performance, we evaluate different possibilities. First, we estimate $\mu_{neu}^{f^s}$ and $\sigma_{neu}^{f^s}$ using all the training data (IFN_{Tr}). Following the ideas behind the ideal normalization approach, we also initialize the parameters using only the neutral samples of the training data (IFN_{Tr^N}). Likewise, we estimate the initial parameters with all the test data, including emotional and neutral samples (IFN_{Te}). Finally, we calculate the parameters using the neutral samples of the test data (IFN_{Te^N}). Notice that the IFN_{Te^N} setting initializes the parameters with the ones used for ideal normalization. This initialization is implemented for comparison purpose, since it is not suitable for real applications (the emotional labels are unknown).

4 Database and Acoustic Features

4.1 IEMOCAP Database

The proposed approach is evaluated with the *interactive emotional dyadic motion capture* (IEMOCAP) corpus [37]. Ten trained actors (5 male and 5 female) took part in five dyadic interactions using scripts and spontaneous improvisations, which were carefully selected to elicit happiness, anger, sadness, and frustration. As a result of the spontaneous dialog between the subjects, other emotions were also observed. The spontaneous interactions between actors elicit expressive reactions with mixed and ambiguous emotions that are similar to the ones observed in naturalistic human interactions (i.e., non-acted data). This is an important characteristic of this corpus that differentiates it from other acted databases in which actors are asked to read sentences portraying a given emotion.

The corpus contains approximately twelve hours of data, which was manually segmented and transcribed at the turn level. We evaluate the emotional content of the corpus using perceptual evaluations by external observers. While it is not clear that perceived emotions match the actual felt emotions [38], emotional labels derived from subjective evaluations are commonly accepted as good approximation of the intended emotion conveyed by the speaker. Each turn was annotated by three evaluators with the following categorical labels: anger, sadness, happiness, disgust, fear, surprise, frustration, excited, neutral and other. An analysis of the assigned labels reveal fair/ moderate agreement across evaluators – details are given in Busso et al. [37]. We only consider samples that reached agreement using majority of votes. Since we are interested in emotion detection (i.e., neutral versus emotional speech), we discard emotional samples that receive neutral labels. We implement this approach to increase the inter-evaluator agreement, reducing the ambiguity in the emotional labels.

4.2 Acoustic Features and Feature Selection

The study considers the set of acoustic features proposed for the INTERSPEECH 2011 Speaker State Challenge [39]. The set includes an exhaustive number of sentence level features that has been commonly used for emotion recognition. First, 59 frame-by-frame features are extracted from each sentence including prosodic (e.g., energy and F0 contour), spectral (e.g., MFCCs, RASTA, spectral flux, and spectral entropy) and voice quality features (e.g., jitter, shimmer). Table 2 lists these *low level descriptors* (LLDs). Then, functionals such as mean, maximum, range, kurtosis, skewness and quartiles are extracted for each of the frame-by-frame features. Table 3 reports these functionals. Altogether, the set includes 4368 sentence level features, which are extracted using the openSMILE toolkit [40].

Given the high dimension of the feature space, we reduce the set using feature selection. Instead of using a wrapper method, in which the performance of a classifier is used as a criterion, we implement a *correlation feature selection* (CFS) technique [41]. CFS aims to identify a feature set that has low correlation between the selected features, but high correlation between the features and the emotional labels. We implement CFS with best first search approach, which is a greedy hill-climbing search method with backtracking capability. Starting from an empty set, the best first search method sequentially adds feature

TABLE 2
The set of frame-level acoustic features used in this study. This set is referred to as *low level descriptors* (LLDs) in the Interspeech 2011 speaker state challenge [39].

Spectral Features
Spectral energy 25-650Hz, 1k-4kHz
Spectral roll-off point 0.25, 0.50, 0.75, 0.90
Spectral flux, entropy, variance, skewness, kurtosis, slope
Rasta
RASTA-style filtered auditory spectrum bands 1-26 (0-8kHz)
MFCC
MFCCs 1-12
Energy
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-crossing rate
Fundamental frequency (F0)
F0
Probability of voicing
Voice Quality (VQ)
Jitter (local, delta)
Shimmer (local)

TABLE 3
The set of sentence-level functionals extracted from the LLDs (see Table 2).

33 base functionals
Quartiles 1-3
3 inter-quartile ranges
1% percentile (≈min), 99% percentile (≈max)
Percentile range 1%-99%
Arithmetic mean, standard deviation
Skewness, kurtosis
Mean of peak distances
Standard deviation of peak distances
Mean value of peaks
Mean value of peaks-arithmetic mean
Linear regression slope and quadratic error
Quadratic regression a and b and quadratic error
Contour centroid
Duration signal is below 25% range
Duration signal is above 90% range
Duration signal is rising/falling
Gain of linear prediction (LP)
LP coefficients 1-5
6 F0 functionals
Percentage of non-zero frames
Mean, max, min, standard deviation of segments length
Input duration in seconds

to the current subset and grades the subset using correlation base criterion. If five consecutive features do not improve the performance of the feature set, the feature selection stops. This strategy helps the search to avoid local maxima. Although a wrapper feature selection approach may provide a more discriminative feature set, we select CFS to fix the feature set for all the experiments. Since this feature selection approach does not depend on any particular classifier, the reported results are not biased to any condition and can be directly compared. Notice that feature selection is conducted after ideal normalization (see Sec 3.1).

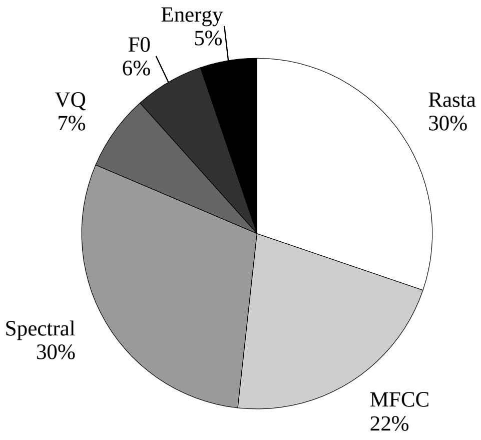


Fig. 4. Distribution of the sentence level features, or functional, selected by CFS per feature group listed in Table 2.

For emotion detection, CFS selected 172 features out of 4368 sentence level features. Figure 4 shows the distribution of the selected features across the six most important LLD categories listed in Table 2.

5 Experiments and Results

We have used a preliminary version of the proposed front-end scheme showing promising results [8], [9], [42], [43]. For example, the normalization scheme was part of the framework that was awarded the first place in the *Intoxication Sub-Challenge* at Interspeech 2011 [42], [43]. This section provides an extensive evaluation of the proposed front-end scheme. The evaluation includes emotion detection problems (Sec. 5.1, neutral versus emotional speech), convergence analysis (Sec. 5.2), sensitivity analysis against errors on speaker identity (Sec. 5.3) and performance in recordings obtained under unconstrained experimental conditions (Sec. 5.4).

5.1 Emotionally-Expressive Speech Detection

The first experiment consists in evaluating the effect of the IFN approach in emotion detection problems – binary classification between neutral and emotional speech. The sentences with emotional labels are clustered together, forming two classes (neutral versus emotional). This scheme yields 1125 neutral and 3839 emotional utterances. The evaluation considers classification (a) *without normalization* (WN), (b) with *global normalization* (GN), (c) with *ideal normalization* (IN) (Sec. 3.1), and (d) with the IFN approach (Sec. 3.2). In global normalization, the mean and standard deviation of the features are estimated across the entire training or testing data of the speakers (speaker-independent normalization). Ideal normalization corresponds to the speaker-dependent normalization scheme, estimated from the neutral portions of each speaker (Sec 3.1). For consistency, all

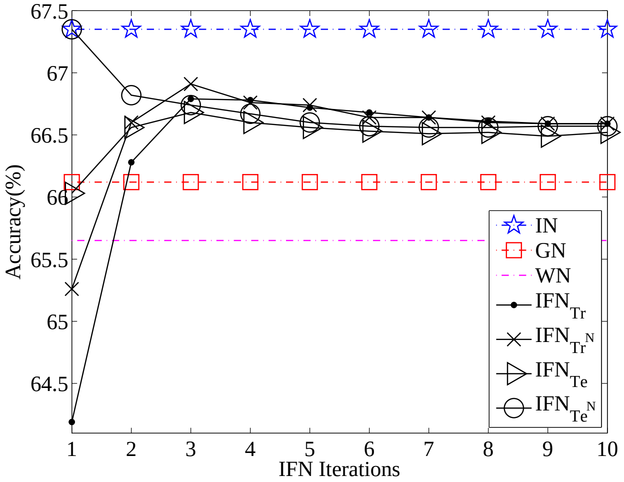


Fig. 5. Performance of the emotional speech detection system with different feature normalization condition. Results are reported in terms of average accuracy across all the speakers. IN: ideal normalization, GN: global normalization, WN: without any feature normalization, IFN: iterative feature normalization. Normalization parameters of the IFN approach are initialized with IFN_{Tr}: all the training set, IFN_{Tr^N}: the neutral samples of the training set, IFN_{Te}: all the testing set, IFN_{Te^N}: the neutral samples of the testing set.

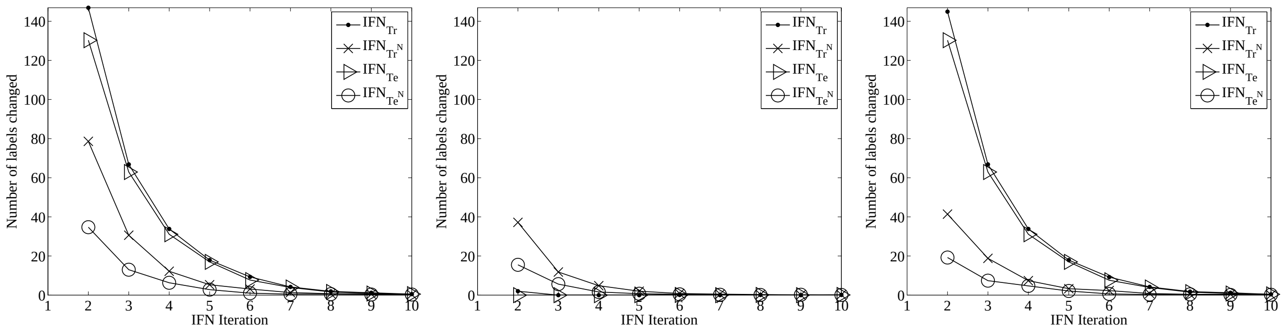
the experiments are conducted with linear kernel SVM classifiers trained with SMO. The evaluation is implemented with leave-one-speaker-out, ten-fold cross validation approach (i.e., speaker-independent evaluation). To avoid dealing with unbalanced classes for the experiments, the emotional utterances are randomly selected to match the number of neutral sentences for both training and testing sets (i.e., training with under-sampling). To make use of the entire database, this random selection is repeated ten times. Hence, the results correspond to the average results achieved across all random selections of the ten folds.

Figure 5 gives the performance of the classifiers trained with different normalization strategies. The results are presented in terms of accuracy. With this metric, the chance level is 50%. The best performance is achieved with the ideal feature normalization approach described in Section 3.1. This result confirms the intuition behind this feature normalization (Fig. 2). The figure shows that global normalization can also improves the emotion discrimination. However it is not as effective as the ideal normalization. The SVM implementation of WEKA provides a prediction probability per each emotional class. This probability is used to select the neutral samples for the IFN parameter estimations. Our preliminary analysis indicated that selecting samples with neutral probability of 0.3 or higher yields better performance than 0.5. Notice that this lower threshold increases the number of samples used to estimate the neutral parameters. Therefore, it gives a more robust estimation with lower variance. The accuracies across iterations are given in Figure 5 for each of the four initialization approaches described in Section 3.3. Notice that the first iteration of IFN_{Te^N} corresponds to the ideal normalization. However, after the first iteration the errors in detecting the neutral and emotional labels causes a deviation from the optimal parameters. Therefore, the performance decreases for the IFN_{Te^N} setting. Ten iterations are used as the stopping criterion. However, after few iterations, different initializations of the IFN converge almost to the same accuracy level. According to the proportion hypothesis test, the differences in performance across different initialization schemes are not statistically significant after 10 iterations ($p-value > 0.46$). This is an important finding since it indicates the stability of the proposed method regardless of the initial parameter estimation. The IFN approach improves the performance of the classifiers, compared to the cases with global normalization and without normalization. Notice, that the ideal normalization is the upper bound performance of the IFN approach (case where the emotion detection system does not have errors). According to the proportion hypothesis test, the effect of ideal normalization is statistically significant compared to the cases without normalization ($p-value < 1e-20$) or with global normalization ($p-value < 0.003$). Also, the IFN approach gives statistically significant improvement over the classifier trained without feature normalization ($p-value < 0.026$)

5.2 Convergence of IFN

Figure 6 shows the number of emotional labels (neutral versus emotional) changed by the emotion detection system in each iteration of the IFN approach. This figures shows that after five iterations all instances of the IFN approach converge, regardless of the initialization process. When the parameters are initialized with neutral samples (IFN_{Tr^N}, IFN_{Te^N}), the number of samples changing labels from emotional to neutral is almost the same as the number of samples changing from neutral to emotional. However, when we initialize the parameters with all the samples (IFN_{Tr}, IFN_{Te}), we observe more samples changing from neutral to emotional classes than from emotional to neutral classes. During training, the features are normalized by statistics from neutral samples. However, the initialization of the normalization parameters during testing is done with all the samples. Therefore, the emotion detection systems yield higher number of false neutral detections. On average, the number of samples that changed labels after five iterations is less than 11 samples, which only represents 0.48% of the data. This criterion can be used as a stopping threshold.

In each iteration, the IFN tries to improve the estimate of the normalization parameters, converging to



(a) Labels that change across iteration (b) Emotional labels that change (c) neutral labels that change

Fig. 6. Number of labels changed in the test data for different iterations of the IFN approach. The normalization parameters of the IFN method are initialized with IFN_{Tr}: all the training set, IFN_{Tr^N}: the neutral samples of the training set, IFN_{Te}: all the testing set, and IFN_{Te^N}: the neutral samples of the testing set. (a) all samples that changed labeled, (b) samples changed from emotional to neutral, (c) samples changed from neutral to emotional.

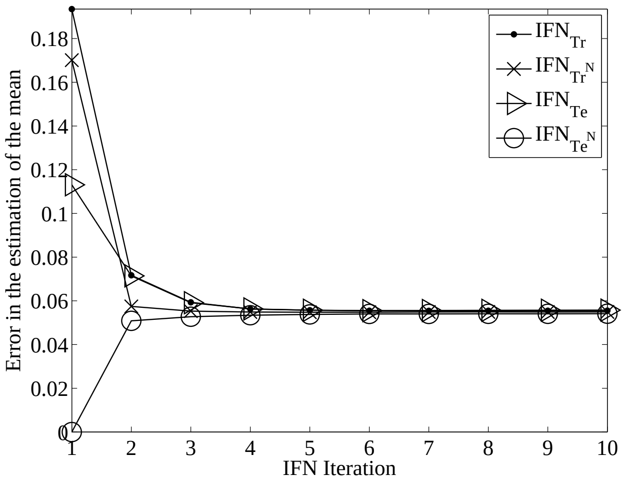


Fig. 7. Estimation error of the mean in the z-standardization for different iterations. The normalization parameters of the IFN approach are initialized with IFN_{Tr}: all the training set, IFN_{Tr^N}: the neutral samples of the training set, IFN_{Te}: all the testing set, and IFN_{Te^N}: the neutral samples of the testing set.

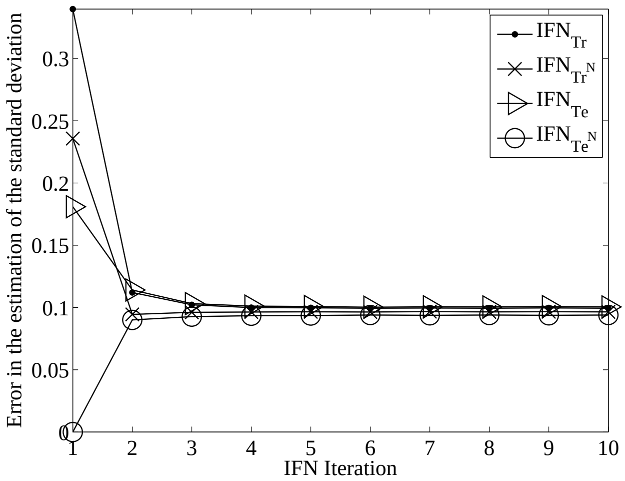


Fig. 8. Estimation error of the standard deviation in the z-standardization for different iterations. The normalization parameters of the IFN method are initialized with IFN_{Tr}: all the training set, IFN_{Tr^N}: the neutral samples of the training set, IFN_{Te}: all the testing set, and IFN_{Te^N}: the neutral samples of the testing set.

a sub-optimal set of normalization parameters. These sets are pretty close to the optimal normalization parameters (i.e., ideal normalization). Figures 7 and 8 show the average absolute errors in the estimation of $\mu_{neu}^{f^s}$ and $\sigma_{neu}^{f^s}$ across the entire selected feature set, with respect to the optimal parameters corresponding to the ideal normalization. Notice that the features have different dynamic range. For better visualization, we have normalized the errors by dividing them by the corresponding standard deviation of each feature. This normalization approach has the same effect as using z-normalization before running the experiments. Notice that the initial step of IFN_{Te} corresponds to the ideal normalization, which yields zero estimation errors. In the following iterations, the emotion detection errors increase the errors of the estimated parameters. For other initialization settings, the error decreases during the IFN iterations. These results indicate the stability of the proposed IFN method.

5.3 Performance with Speaker Identification Error

The proposed IFN approach relies on the speaker identity given for the test data. We assume that test samples come from a single speaker. This section studies the effect of speaker identification errors in the test set. The data is divided into two sets of five speakers each. The emotion detection models are built on one set by using the correct speaker identity. Then, the data of each of the test speakers is mixed with the other four test speakers according to the target percentage in speaker errors. This

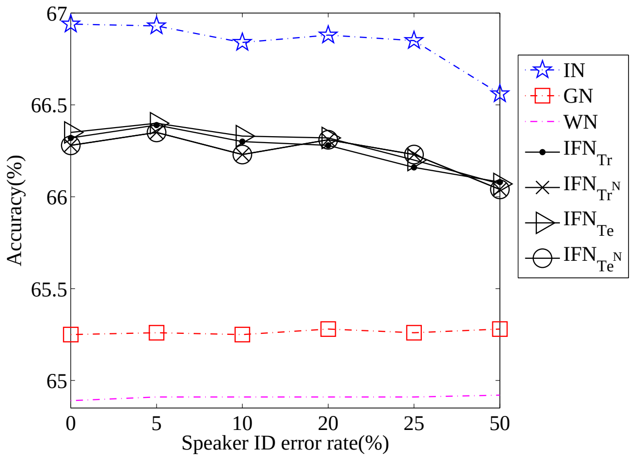


Fig. 9. Accuracy (%) of the emotion detection system with errors in speaker identification.

scheme simulates the errors introduced by a speaker identification system. Notice that the parameters of IN are estimated with all the neutral samples associated with each speaker, which may contain neutral samples of the other speakers as well. Given the differences in experimental settings, the results of this section are different from the ones presented in Section 5.1, even when no error in speaker identity is introduced (classifiers are trained with only 5 speakers). Similar to Section 5.1, the training and testing partitions are randomly balanced. Different error rates in speaker identification are introduced among the samples in the testing partition (e.g., 0%, 5%, 10%, 20%, 25% and 50%). Then, the different normalization schemes are evaluated under different error rates on speaker identification.

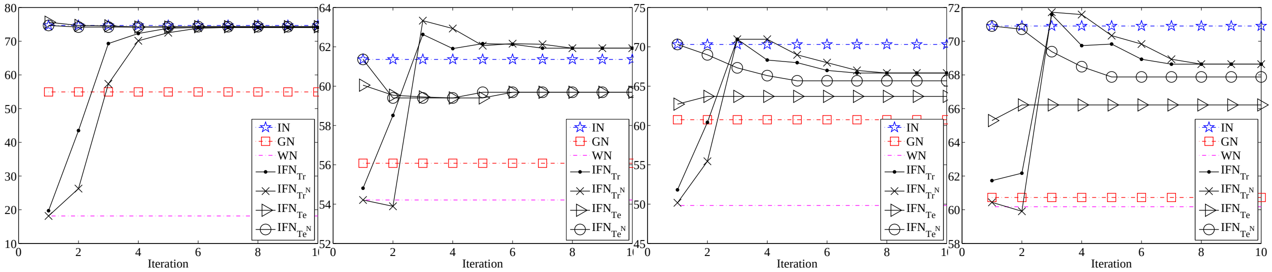
The reported results correspond to the average obtained by considering each of the two sets as training and testing partitions (two-fold cross-validation). Figure 9 displays the achieved average performance for each of the normalization schemes. Since the condition without feature normalization (WN) and global normalization (GN) perform the classifications without any assumption on the speaker identity, the accuracy is constant regardless of the percentage of speaker identity error parameter. However, the performance of ideal normalization drops when the error rate of the speaker identity increases. IFN, as a speaker-dependent front-end, normalizes the feature space to better match the feature space in the training data. According to Figure 9, IFN presents robust performance against speaker identity errors, dropping the accuracy only when speaker identity error is 50%. However, it still outperforms the global normalization setting. This result is important since, if needed, the accuracy of a speaker identification system does not need to be perfect for the IFN system to provide accurate emotion detection results.

5.4 Performance in Uncontrolled Recordings

The proposed normalization approach is finally evaluated on data from realistic recordings in uncontrolled settings. For this purpose, we downloaded from a video-sharing website several talks and interviews given by a recognized celebrity (only the audio is used in this study). The videos span different ages of the target individuals (from 15 to 30 years), and were recorded in different environmental conditions. They include various undesirable factors such as background music, background noise and overlapped speech of multiple speakers (the recordings come from interviews during TV shows). Dealing with all these factors is a challenging problem that requires insights from speech enhancement, voice activity detection, speaker diarization, among other related fields. Therefore, we manually remove noisy and overlapped segments to simplify the evaluation of the IFN approach in naturalistic recordings.

The corpus was split into 5 sec segments, which were emotionally annotated by six graduate students. Unlike similar studies that consider emotional dimensions such as valence and arousal [44], [45], we directly asked the evaluators to assess whether the samples are emotional or neutral. The subjects used a slider bar to assess each speech segment on a continuous scale from 0 (neutral) to 1 (emotional). To estimate the reliability of the perceptual evaluation, we estimated the correlation between each evaluator and the average scores across the other five evaluators. In average, we observe a correlation of ρ=0.52, which represents a strong agreement for this task. Notice that this low level of agreement is commonly observed across perceptual emotional evaluations. The average of the scores across evaluators is considered as ground truth. The assigned scores are skewed toward 0 (neutral samples). This result is expected, since most of the sentences do not convey emotional information in natural recordings. To address the unbalanced emotional content of the corpus, the speech segment is considered as *neutral* if its average value is lower than 0.4. Otherwise, it is considered as *emotional*. This threshold is similar to the ones used in previous studies [46], [47]. Altogether, the corpus includes 837 speech files, with more neutral (727) than emotional (110) samples. We evaluate the IFN approach with unbalanced and balanced classes in the testing set (see Fig. 10). Detailed information about this corpus is given by Rahman and Busso [9].

The uncontrolled recordings are only used for testing the IFN approach. The training of the emotion detection system is implemented with the IEMOCAP database. Similar to the experiments presented in Section 5.1, we select a balanced set of neutral and emotional sentences for each of the ten speakers in the IEMOCAP corpus. Then, the SVM-based emotion detection models are built with this data. Given that



(a) unbalanced data - accuracy (b) unbalanced data - F-score (c) balanced data - accuracy (d) balanced data - F-score

Fig. 10. Performance of emotion detection methods in uncontrolled recordings for unbalanced and balanced settings (*WN*= without normalization,*GN*= global normalization *IFN*= iterative feature normalization, and *IN*= ideal normalization).

we present results with balanced and unbalanced classes in the testing partitions, we measure the performance with F-score (F), in addition to *accuracy* (A). Equation 2 gives the formula for F, where $\bar{R}$ and $\bar{P}$ are the average recall and average precision of the two classes, respectively. Notice that we estimate both the recall and precision rate for the neutral and emotional classes, separately, and then we take their corresponding averages. Therefore, F is not sensitive to the selection of the positive class. A random classifier gives an F-score of 50% regardless of the prior distribution of the classes.

$$F = 2\frac{\bar{P}\bar{R}}{\bar{P}+\bar{R}} \tag{2}$$

5.4.1 Evaluation with unbalanced classes

Figures 10(a-b) give the evaluations on the uncontrolled recordings when the classes in the testing set are highly unbalanced (110 versus 727). These figures give the performance using the same initialization conditions described in section 5.1. They also provide the performance of the emotion detection system with global feature normalization and without feature normalization. According to the proportion hypothesis test, the F-scores reported in Figure 10(b) are significantly higher than 50% ($p-value < 1e-20$). Figure 10 shows that when no normalization approach is used, the emotion detection system cannot cope with the mismatched condition of training with the IEMOCAP corpus and testing with the uncontrolled recordings (18.2% accuracy). Similarly, the global normalization does not yield satisfactory performance (55% accuracy). However, the ideal normalization provides more robust results. Notice that accuracy of ideal normalization is more than 19% higher than global normalization. In spite of the challenging recording conditions and the highly unbalanced corpus, the system trained with the IFN approach is also able to provide results comparable to the ideal normalization. These results suggest the potential of the IFN approach on real-life, non-laboratory data.

5.4.2 Evaluation with balanced classes

We also evaluate the performance of the IFN approach when the emotional classes are balanced in the test set. We select the top 111 neutral samples with the lowest ratings to achieve a fairly balanced set (110 emotional, 111 neutral). Figures 10(c-d) show the accuracy and F-score achieved by the corresponding emotion detection systems using the balanced dataset. The figures show that the IFN approach outperforms global normalization. It also provides better performance than when no feature normalization is used. The figures reveal trends consistent with the rest of the evaluation, highlighting the benefits of using the IFN approach.

6 Discussion and Conclusions

This paper presented the *iterative feature normalization* (IFN) framework as an unsupervised front-end for emotion recognition systems. The speaker-dependent approach iteratively detects emotionally neutral samples which are used to estimate the normalization parameters. The normalization is applied to the entire corpus, including the emotional samples. This normalization reduces the differences in neutral speech across speakers and recordings, while preserving the differences between emotional classes in the feature space. An exhaustive evaluation is conducted to assess the performance of the IFN approach. The results reveal that the performance of the emotion detection (neutral versus emotional speech) based on the IFN framework gives better accuracies than the ones achieved with classifiers trained without normalization or with global normalization. The IFN approach also improves the accuracy in detecting emotional speech obtained from real life, unconstrained recordings. While the approach is speaker-dependent, the evaluations reveal that the performance is not very sensitive to speaker identification errors, which make it suitable for practical applications.

There are several interesting directions that we are considering to improve the proposed IFN approach. First, the current implementation based on z-standardization corresponds to a simple affine transformation. We are exploring the benefits of using other transformations including feature warping. Another research direction is developing approaches to recognize the identity of the speakers, which is a non trivial task with emotional speech. We are exploring speaker clustering strategies that will reduce the impact of speaker identification errors. We also leave as future work the evaluation of the IFN approach with overlapped speech collected in noisy environments during multiparty interactions. These challenging conditions will affect the estimation of the speaker-dependent normalization parameters. Potential solutions are to detect overlapped speech, so that these segments can be discarded, and to apply speech enhancement solutions. Finally, this unsupervised front-end framework can be coupled with model adaptation strategies.

Acknowledgments

This study was funded by National Science Foundation (NSF) grant IIS 1217104 and Samsung Telecommunications America as well as grants to USC SAIL from NSF (IIS-0911009, CCF-1029373) and DoD.

References

[1] O. Küstner, R. Tato, T. Kemp, and B. Meffert, "Towards real life applications in emotion recognition," in *Affective Dialogue Systems(ADS 2005), Lecture Notes in Artificial Intelligence 3068*, E. André, L. Dybkaer, W. Minker, and P. Heisterkamp, Eds. Berlin, Germany: Springer-Verlag Press, May 2004, pp. 25–35.

[2] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.

[3] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.

[4] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, June 2008.

[5] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The prosody module," in *VERBMOBIL: Foundations of Speech-to-speech Translations*, M. Maybury, O. Stock, and W. Wahlster, Eds. Berlin, Germany: Springer Verlag, 2000, pp. 106–121.

[6] C. Busso, S. Lee, and 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.

[7] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 320–323.

[8] C. Busso, A. Metallinou, and S. Narayanan, "Iterative feature normalization for emotional speech detection," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 5692–5695.

[9] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

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

[11] 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.

[12] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*. Piscataway, NJ, USA: IEEE Press, 2000.

[13] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and 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.

[14] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003)*, vol. 2, Hong Kong, China, April 2003, pp. 1–4.

[15] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, November-December 2011.

[16] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 597–600.

[17] A. Metallinou, A. Katsamanis, and S. Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 2401–2404.

[18] T.-L. Pao, J.-H. Yeh, Y.-T. Chen, Y.-M. Cheng, and Y.-Y. Lin, "A comparative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, D.-S. Huang, L. Heutte, and M. Loog, Eds. Berlin, Germany: Springer-Verlag Press, July 2007, pp. 997–1005.

[19] T.-L. Pao, C. Chien, Y.-T. Chen, J.-H. Yeh, Y.-M. Cheng, and W.-Y. Liao, "Combination of multiple classifiers for improving emotion recognition in Mandarin speech," in *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2007)*, vol. 1, Kaohsiung City, Taiwan, November 2007, pp. 35–38.

[20] Z. Yan, Z. Li, Z. Cairong, and Y. Yinhua, "Speech emotion recognition using modified quadratic discrimination function," *Journal of Electronics (China)*, vol. 25, no. 6, pp. 840–844, November 2008.

[21] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. Picard, Eds. Berlin, Germany: Springer Berlin / Heidelberg, September 2007, pp. 139–147.

[22] B. Schuller, B. Vlasenko, R. Minguez, G. Rigoll, and A. Wendemuth, "Comparing one and two-stage acoustic modeling in the recognition of emotion in speech," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2007)*, Kyoto, Japan, December 2007, pp. 596–600.

[23] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, pp. 613–625, July-August 2010.

[24] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, pp. 119–131, July-Dec 2010.

[25] X. Le, G. Quénot, and E. Castelli, "Recognizing emotions for the audio-visual document indexing," in *Ninth International Symposium on Computers and Communications (ISCC 2004)*, vol. 2, Alexandria, Egypt, June-July 2004, pp. 580–584.

[26] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," in *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerland, September 2003, pp. 125–128.

[27] T. Zhang, M. Hasegawa-Johnson, and S. Levinson, "Mental state detection of dialogue system users via spoken language," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR-2003)*, Tokyo, Japan, April 2003.

[28] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker normalisation for speech based emotion detection," in *15th International Conference on Digital Signal Processing (DSP 2007)*, Cardiff, Wales, UK, July 2007, pp. 611–614.

[29] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. In Press, 2013.

[30] Z. Zeng, J. Tu, B. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T. Huang, and S. Levinson, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, vol. 2, San Diego, CA, USA, June 2005, pp. 967–972.

[31] 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)*, vol. 2, Wuhan, China, December 2008, pp. 140–144.

[32] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory (COLT 1998)*, Madison, WI, USA, July 1998, pp. 92–100.

[33] C. Busso, P. Georgiou, and S. Narayanan, "Real-time monitoring of participants interaction in a meeting using audio-visual sensors," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, vol. 2, Honolulu, HI, USA, April 2007, pp. 685–688.

[34] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*. State College, PA: ACM Press, October 2004, pp. 205–211.

[35] T. Rahman, S. Mariooryad, S. Keshavamurthy, G. Liu, J. Hansen, and C. Busso, "Detecting sleepiness by fusing classifiers trained with novel acoustic features," in *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, August 2011, pp. 3285–3288.

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

[37] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[38] C. Busso and S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 257–260.

[39] 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.

[40] 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.

[41] M. A. Hall, "Correlation based feature-selection for machine learning," Ph.D. dissertation, The University of Waikato, Hamilton, New Zealand, April 1999.

[42] D. Bone, M. P. Black, M. Li, A. Metallinou, S. Lee, and S. Narayanan, "Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011, pp. 3217–3220.

[43] D. Bone, M. Li, M. Black, and S. Narayanan, "Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors," *Computer, Speech, and Language*, October 2012.

[44] 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: ISCA, September 2000, pp. 19–24.

[45] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.

[46] 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.

[47] J. Arias, C. Busso, and N. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech and Language*, vol. In Press, 2013.

Carlos Busso (S02-M09) is an Assistant Professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He received his B.S (2000) and M.S (2003) degrees with high honors in electrical engineering from University of Chile, Santiago, Chile, and his Ph.D (2008) in electrical engineering from University of Southern California (USC), Los Angeles, USA. He was selected by the School of Engineering of Chile as the best Electrical Engineer graduated in 2003 across Chilean universities. At USC, he received a Provost Doctoral Fellowship from 2003 to 2005 and a Fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests are in digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human interaction, in-vehicle active safety system, and machine learning methods for multimodal processing.

Soroosh Mariooryad (S'2012) received his B.S degree (2007) with high honors in computer engineering from Ferdowsi University of Mashhad, and his M.S degree (2010) in computer engineering (artificial intelligence) from Sharif University of Technology (SUT), Tehran, Iran. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas (UTD), Richardson, Texas, USA. From 2008 to 2010, he was a member of the Speech Processing Lab (SPL) at SUT. In 2010, he joined as a research assistant the Multimodal Signal Processing (MSP) laboratory at UTD. In summer 2013, he interned at Microsoft Research working on analyzing speaking style characterstics. His research interests are in speech and video signal processing, probabilistic graphical models and multimodal interfaces. His current research includes modeling and analyzing human non-verbal behaviors, with applications to speech-driven facial animations and emotion recognition. He has also worked on statistical speech enhancement and fingerprint recognition.

Angeliki Metallinou received her Diploma in Electrical and Computer Engineering from the National Technical University of Athens, Greece, in 2007, and her Masters and PhD degrees in Electrical Engineering in 2009 and 2013 respectively, from University of Southern California (USC). Between 2007 and 2013 she has been a member of the Signal Analysis and Interpretation Lab (SAIL) at USC, working on spoken and multimodal emotion recognition and computational approaches for healthcare. During summer 2012, she interned at Microsoft Research working on spoken dialog systems. She is currently working as a research scientist at Pearson Knowledge Technologies, on automatic speech recognition and language assessment for education applications, and on remote healthcare monitoring. Her research interests include speech and multimodal signal processing, affective computing, machine learning and dialog systems.

Shrikanth (Shri) Narayanan (StM'88-M'95-SM'02-F'09) is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics and Psychology and as the founding director of the Ming Hsieh Institute. Prior to USC he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered information processing and communication technologies with a special emphasis on behavioral signal processing and informatics. [http://sail.usc.edu]

Prof. Narayanan is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an Editor for the Computer Speech and Language Journal and an Associate Editor for the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, APSIPA TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING and the Journal of the Acoustical Society of America. He was also previously an Associate Editor of the IEEE TRANSACTIONS OF SPEECH AND AUDIO PROCESSING (2000-2004), IEEE SIGNAL PROCESSING-MAGAZINE (2005-2008) and the IEEE TRANSACTIONS ON MULTIMEDIA (2008-2011). He is a recipient of a number of honors including Best Transactions Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010-2011. Papers co-authored with his students have won awards at Interspeech 2013 Paralinguistics Challenge, Interspeech 2012 Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, InterSpeech 2010, InterSpeech 2009-Emotion Challenge, IEEE DCOSS 2009, IEEE MMSP 2007, IEEE MMSP 2006, ICASSP 2005 and ICSLP 2002. He has published over 500 papers and has fourteen granted U.S. patents.