Compensating for Speaker or Lexical Variabilities in Speech for Emotion Recognition

Soroosh Mariooryad and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Electrical Engineering Department,
The University of Texas at Dallas, 800 West Campbell Road, Richardson, TX, USA

Abstract

Affect recognition is a crucial requirement for future human machine interfaces to effectively respond to nonverbal behaviors of the user. Speech emotion recognition systems analyze acoustic features to deduce the speaker's emotional state. However, human voice conveys a mixture of information including speaker, lexical, cultural, physiological and emotional traits. The presence of these communication aspects introduces variabilities that affect the performance of an emotion recognition system. Therefore, building robust emotional models requires careful considerations to compensate for the effect of these variabilities. This study aims to factorize speaker characteristics, verbal content and expressive behaviors in various acoustic features. The factorization technique consists in building phoneme level trajectory models for the features. We propose a metric to quantify the dependency between acoustic features and communication traits (i.e., speaker, lexical and emotional factors). This metric, which is motivated by the mutual information framework, estimates the uncertainty reduction in the trajectory models when a given trait is considered. The analysis provides important insights on the dependency between the features and the aforementioned factors. Motivated by these results, we propose a feature normalization technique based on the whitening transformation that aims to compensate for speaker and lexical variabilities. The benefit of employing this normalization scheme is validated with the presented factor analysis method. The emotion recognition experiments show that the normalization approach can attenuate the variability imposed by the verbal content and speaker identity, yielding 4.1% and 2.4% relative performance improvements on a selected set of features, respectively.

Keywords: emotion recognition, factor analysis, feature normalization, speaker variability

1. Introduction

Expressing and perceiving emotions play fundamental roles in human interaction. A message can be interpreted differently depending on its intrinsic affective content. Even our decisions, at a cognitive level, are influenced by emotions (Hastie and Dawes, 2001). Therefore, emotions can clarify ambiguities in the message and affect the listeners' reactions. Hence, it is important for a *human machine interface* (HMI) to recognize the affective state of the user and respond accordingly. Human voice is a rich communicative channel to externalize emotions. However, this channel conveys other communicative aspects, such as the lexical content and the idiosyncratic characteristics of the speaker. All these factors increase the variability observed in the acoustic signal. A robust emotion recognition system should compensate for the underlying lexical- and speaker-dependent variabilities. By reducing these variabilities, the emotion classifiers should provide better predictions of the users' affective states. In this context, this paper aims to quantify the speaker, lexical and emotional dependencies on various frame level acoustic features. Building upon the analysis, the study proposes a normalization scheme to mitigate the variabilities caused by the speaker dependency and underlying lexical content in the context of emotion recognition.

While speaker normalization schemes have been proposed for emotion recognition (Busso et al., 2011; Vlasenko et al., 2007; Wöllmer et al., 2008; Schuller et al., 2010), there are few approaches that have attempted to compensate for lexical dependency (Fu et al., 2008). The externalization of emotion is not uniformly conveyed across time. This nonuniform encoding of information has been observed at various linguistic levels such as sentence, phrase, word and phoneme (Yildirim et al., 2004; Lee et al., 2005; Busso et al., 2007; Busso and Narayanan, 2006, 2007; Vlasenko et al., 2011b). These studies reveal an interplay between the externalization of emotion and speech production (Busso and Narayanan, 2007). As a result, certain linguistic units are more affected by emotion modulation, which should be properly modeled in a robust emotion recognition system. With a given set of standard features, compensation schemes can be implemented at feature (Busso et al., 2011; Rahman and Busso, 2012) or model (Lee et al., 2004) level. While both approaches have advantages and limitations, this study focuses on the front-end of an emotion recognition system.

This study proposes a framework to identify the lexical, speaker and emotional dependencies on the acoustic features. The approach, which was originally developed to localize expressive areas in the face (Mariooryad and Busso, 2012), consists in measuring the decrease in uncertainty when the state of one of the factors (i.e., speaker, lexical or emotion) is known. Separate trajectory models at the phoneme level are used to capture the temporal pattern of various acoustic features. The feature trajectory is represented by a finite dimensional vector after resampling and interpolating the signal. Then, the sample

covariance matrix is estimated over multiple repetitions of the phonemes. Once the trajectory models are built, the uncertainty is measured in term of the trace of the covariance matrix (i.e., sum of the eigenvalues). For each acoustic feature, the proposed factor analysis identifies the dominant factor yielding the highest uncertainty reduction.

Motivated by the trajectory models, a normalization scheme based on the whitening transformation is proposed to compensate for the speaker and lexical variabilities. This scheme is developed based on the lexical normalization approach previously introduced to enhance the performance of facial expressions recognition systems (Mariooryad and Busso, 2013b), which is adapted for speech features. The proposed lexical normalization approach assumes that the transcription is available. According to the factor analysis metric, after speaker normalization, the speaker dependency is attenuated in all acoustic features. Therefore, the emotion- and lexical-dependent aspects become the dominant factors after the normalization. Similarly, lexical normalization reduces the effect of the underlying phonetic content in all the acoustic features.

To validate the findings from the analysis, an emotion recognition system is built using the aforementioned normalization schemes. When the data is properly normalized to compensate for the lexical and speaker variabilities, the classifiers achieve 4.1% and 2.4% relative improvement over the baseline performance, respectively. These classifiers are trained with a carefully selected set of features identified by the factor analysis (reduction in 38% of the features). This reduced set is created by identifying the acoustic features that are mainly affected by emotions rather than by the speaker or lexical content variabilities. Interestingly, these selected features preserve the performance of the classifiers trained with all the acoustic features. This result suggests that the proposed shape-based analysis can be used as a principled method to reduce the dimension of the feature vector, which is commonly above 4000 in emotion recognition systems (Schuller et al., 2011b). Furthermore, the insights of this study not only can guide the design of robust emotion classification systems, but also inspire solutions in other speech processing tasks such as speech recognition and speaker identification.

The rest of the paper is organized as follows. Section 2 summarizes the related studies and highlights the contributions of this work. Section 3 describes the database and acoustic features considered in this study. Section 4 presents the proposed framework, including the trajectory models, factor analysis and the whitening normalization scheme. Section 5 provides the results of the factor analysis method, before and after the normalizations. This factor analysis yields a set of acoustic features to characterize the emotions. Section 6 demonstrates the discriminative power of the selected features as well as the benefit of the normalization approach. Section 7 concludes the study with discussion and our future directions.

2. Related Works

Compensating for factors that are not related to the task at hand (i.e., nuisance variabilities) is of great interest in all speech processing tasks, including language recognition (Dehak et al., 2011), speaker identification (Kenny et al., 2007) and emotion recognition (Li et al., 2012). The state-of-the-art compensation methods in speech processing tasks find discriminative subspaces that isolate the variable of interest from the nuisance variabilities. One of these techniques is *joint factor analysis* (JFA), which separates speaker and session variabilities for speaker verification (Kenny et al., 2007). Other compensation methods include *nuisance attribute projection* (NAP) and *linear discriminant analysis* (LDA) in the total variability subspace, which is known as the i-vector space (Dehak et al., 2009). These techniques are also applied in paralinguistic recognition tasks. Li et al. (2012) used latent factor analysis to represent the affective states of the speaker. Xia and Liu (2012) proposed the use of i-vector space for emotion recognition. Despite the distinct nature of these tasks, the studies often consider standard features such as *Mel frequency cepstral coefficients* (MFCCs) and *Mel filter banks* (MFBs). However, emotional speech affects various acoustic and prosodic features. Therefore, robust emotion recognition systems often consider big feature vectors characterizing different aspects of the acoustic properties (Schuller et al., 2011a). The aforementioned methods (i.e., latent factor analysis, i-vector) cannot be easily extended to handle high number of low level descriptors, since these methods already produce high dimensional feature spaces even when 13 MFCCs are used. In contrast, we propose a factor analysis, which considers each low level feature independently. This work systematically studies the dependency of different spectral and prosodic features on the three discussed factors in the context of emotion recognition problem. The proposed method identifies relevant features to model each source of variability.

Various studies have proposed feature normalization techniques to reduce the speaker dependency (Busso et al., 2009; Zeng et al., 2008; Busso et al., 2011; Rahman and Busso, 2012). However, conventional emotion classification approaches generally disregard the dependency on the verbal content of the message, which can result in inaccurate assessments. Previous studies have shown the importance of considering the underlying lexical content to characterize paralinguistic information. Hansen and Womack (1996) observed characteristic phoneme-dependent patterns in speech under physical stress. Lee et al. (2004) showed that the magnitude and direction of the formant changes in emotional speech vary across vowels. They clustered the phonemes into broad phonetic classes for emotion recognition experiments. They showed that phoneme-class-dependent *hidden Markov models* (HMMs) outperform generic phoneme-independent HMMs for emotion recognition. In our previous work, we demonstrated that the likelihoods of neutral phoneme-dependent models can be used to discriminate between different emotions (Busso et al., 2007). Vlasenko et al. (2011b,a) analyzed the effect of emotion on vowel formants. Based on their results, they proposed the use of vowel-dependent models for emotion recognition. Chauhan et al. (2011) showed that text-dependent emotion recognition improves the performance compared to text-independent methods. Fu et al. (2008) reported similar results for emotion recognition. However, it is not practical to build a model for every possible

Figure 1: The setting for the recording of the IEMOCAP corpus.

utterance. Therefore, this study proposes the idea of lexical normalization at phoneme level to suppress the variability imposed by verbal contents. The emotion recognition experiments supports the effectiveness of the proposed normalization to achieve this goal. The primary contributions of this paper are:

- An exhaustive analysis of various acoustic features to identify their dependencies at phoneme level on lexical, speaker and emotional aspects (with main focus on emotion recognition).
- A novel trajectory-based normalization scheme to compensate for speaker or lexical variabilities.

3. Database and Acoustic Features

3.1. IEMOCAP Database

The factor analysis and emotion recognition tasks are conducted on the *interactive emotional dyadic motion capture* (IEMOCAP) database (Busso et al., 2008). This corpus was collected to study spontaneous emotional interactions, recorded under controlled conditions with multiple sensor technologies (microphones, camera and motion capture system). Figure 1 depicts the data collection setting. The corpus contains five sessions of dyadic conversations between two trained actors (12 hours of data). In total, ten subjects participated in the recordings (five females and five males). Two elicitation techniques were used to evoke spontaneous emotional reactions. First, the actors played three scripts carefully designed to elicit happiness, anger, sadness and frustration. Then, the actors participated in improvisations of hypothetical scenarios (e.g., lost baggage in an airport, enrolling in the army, and getting married). Notice that the emotions were elicited as dictated by the dialog, producing realistic, natural recordings full of ambiguous and mixed emotions (Mower et al., 2009). The utterances were manually transcribed and segmented into turns. The phoneme and word boundaries were estimated with forced alignment. Subjective evaluations were conducted to assess the emotional content of the utterances into discrete emotional categories. The discrete labels included anger, sadness, happiness, disgust, fear, surprise, frustration, excited, neutral and other. Although the IEMOCAP database includes motion capture sensors to obtain detailed facial expressions, this study only uses speech recorded with high quality shotgun microphones (Schoeps CMIT 5U). Detailed information about the corpus can be found in Busso et al. (2008).

Table 1: Emotion distribution in the IEMOCAP database. Neu: neutral, Hap: happiness, Ang: anger, Sad: sadness

Emotion	Hap	Ang	Sad	Neu	All
Number	950	678	955	1112	3695

A key aspect of the proposed approach is that the acoustic features are represented with trajectory models (see Sec. 4.1). Building these models requires enough samples. Therefore, the study uses only samples labeled with the four most frequent emotional classes in the IEMOCAP database: happiness, anger, sadness and neutral state. Similar to previous works (Metallinou et al., 2010a; Mariooryad and Busso, 2013a), we have merged the samples labeled as excited and happiness, into a single class. The distribution of emotions in the selected set is given in Table 1. Batliner et al. (2010) studied different lexical units that are appropriate for emotion recognition. They proposed word as the smallest adequate lexical unit for the analysis of emotions. However, it is not feasible to create lexicon-dependent model per word, so we consider phoneme as the basic lexical unit. Notice that previous studies have used phoneme-level processing to model emotions (Vlasenko et al., 2011b; Lee et al., 2004). This work considers phoneme as the basic lexical unit for processing. The phonetic alphabet used for forced alignment consists of 47 phones. The study considers only the symbols with more than 100 repetitions per emotion, reducing the set to 41 phones. This group spans 99% of the actual phones comprising the data. For the emotion recognition evaluations, we use a garbage model for the rest of the phonemes (see Sec. 6). Notice that the ten speakers in the database provide enough diversity to analyze the speaker dependency on the features.

3.2. Acoustic Features

This study uses the exhaustive set of frame-by-frame acoustic features provided by Schuller et al. (2011b) for the Interspeech 2011 speaker state challenge (*low level descriptors* (LLDs)). Table 2 summarizes the features into spectral and prosodic features. The auditory spectrum components are obtained after applying the equal loudness curves and loudness compression to the 26 MFB outputs, to mimic human auditory perception (Hermansky, 1990). RASTA-style auditory spectrum components are estimated by temporally filtering the MFBs with *relative spectral* (RASTA) filter and applying the auditory perception filters. The spectral features consist of RASTA-style filtered auditory spectrum components, MFCCs and a set of functions derived from spectral components (i.e., short term Fourier transform). These functions include the energy in low frequency (i.e, fband 25-650Hz) and high frequency (i.e., fband 1k-4kHz) bands, the X roll-off point, which is the frequency

Table 2: Acoustic feature considered in this study. The set corresponds to the *low level descriptors* (LLDs) used for the Interspeech 2011 speaker state challenge (Schuller et al., 2011b).

Spectral related features
RASTA-style filt. auditory spectrum, bands 1-26 (0-8kHz)
MFCC 1-12
Spectral energy 25-650Hz, 1k-4kHZ
Spectral roll off point 0.25, 0.50, 0.75, 0.90
Spectral flux, entropy, variance, skewnewss, kurtosis
Prosodic features
L1 norm of auditory spectrum components (loudness)
L1 norm of RASTA-style filtered auditory spectrum
RMS energy
Zero-crossing rate
F0
Probability of voicing
Jitter (local, delta)

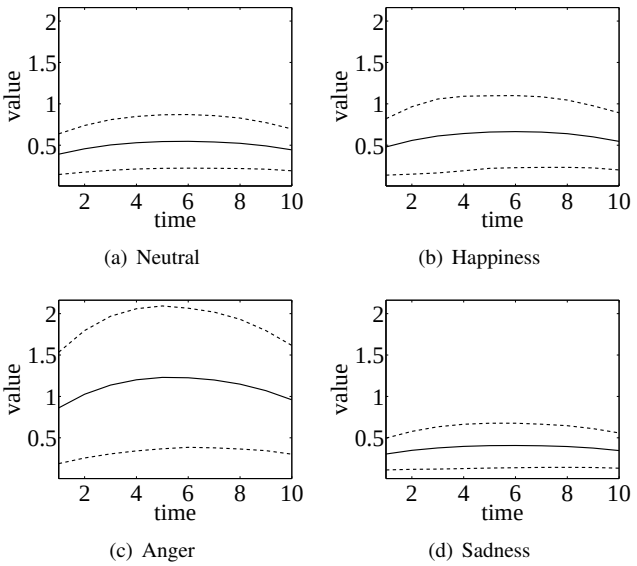

Figure 2: Mean and standard deviation of the trajectories of the *auditory spectrum L1 norm* for phoneme /æ/. For each emotion, the trajectory is extracted across all speakers.

beyond which the total signal energy exceeds the $X * 100\%$ of total signal energy, and other statistics including flux, entropy, variance, skewness and kurtosis. The MFCCs are low pass liftered, which removes the excitation information. The prosodic features include L1 norm of auditory spectrum components, with and without applying the RASTA filters, *root mean square* (RMS) energy, *zero-crossing rate* (ZCR), fundamental frequency (F0), probability of voicing, jitter and its derivative. Jitter is the frame-by-frame pitch period deviations. Note that the F0 contour prediction produces zero values during the unvoiced segments. Therefore, to facilitate the processing, F0 is often interpolated during the unvoiced segments (Chappell and Hansen, 1998; Lei et al., 2006). Given that this study proposes trajectory models to capture the dynamics of the acoustic features, the F0 contour is interpolated with *piecewise cubic Hermite interpolating polynomial* (PCHIP) method. These acoustic features are extracted with the openSMILE toolkit (Eyben et al., 2010). The detailed description of the features can be found in Schuller et al. (2011b).

4. Methodology

To unveil characteristics of the acoustic features, this work adopts the factor analysis model, originally proposed to study facial expressions (Mariooryad and Busso, 2012). In contrast to current studies that consider global statistics from acoustic features, the proposed model directly captures the temporal trajectory of the LLDs. We assume that the phonetic transcription of the corpus is available. This section describes the proposed trajectory model and our approach to quantify uncertainty.

4.1. Trajectory Models for Acoustic Features

Capturing the temporal dynamics of acoustic features is an interesting and novel approach to describe expressive speech. Previous studies have successfully implemented models to capture temporal shape of acoustic features in applications such as word spotting (Gish and Ng, 1993) and automatic speech recognition (Gish and Ng, 1996; Gong, 1997). For each phoneme, the temporal shape of the features is represented by using either a parametric or nonparametric model trained with multiple samples. Gish and Ng (1996) used a second order polynomial model to capture the temporal shape of cepstral coefficients. Gong (1997) proposed a probabilistic trajectory model for this problem.

This study implements a nonparametric approach to capture the temporal dynamics of the acoustic features. First, an interpolation and resampling scheme is used to linearly align all the instances for a given phoneme. This approach generates an N-dimensional vector that describes the shape of the feature. The approach is similar to the preprocessing steps proposed by Gish and Ng (1993, 1996). The length of the vector is empirically set to $N = 10$. Our preliminary experiments showed that increasing the length does not affect the reported results. The average temporal shape for each acoustic feature A is modeled with an $N \times 1$ mean trajectory vector μ_A. An $N \times N$ covariance matrix (Σ_A) is used to captures the variations around this mean trajectory. Figure 2 shows the trajectory models of *auditory spectrum L1 norm* for phoneme /æ/. The trajectories are given for each emotion, across all speakers. The solid line is the mean trajectory (μ_A), and the dashed lines show the standard deviations around this trajectory (i.e., square root of the diagonal elements of Σ_A). This figure not only suggests the existence of consistent lexical pattern, but also highlights the characteristic emotional patterns in the trajectories.

4.2. Measuring Uncertainty and Factor Analysis

The proposed factor analysis approach is based on the premise that conditioning on a relevant factor reduces the entropy of a random variable (Cover and Thomas, 2006). The total variability of a feature is the result of the modulation of

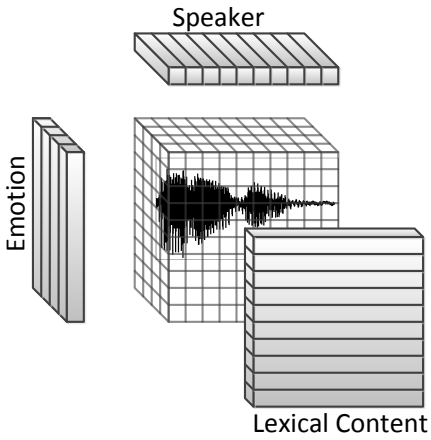

Figure 3: Speaker characteristic, lexical content and emotion are three sources of variability conveyed in the speech signal (the figure does not imply orthogonality between the factors).

all the underlying factors influencing the feature. Therefore, if the space spanned by the feature is conditioned on a relevant factor, the within-class variability is expected to decrease. The amount of variability reduction can be regarded as a measure of dependency between the factor and the feature. The higher the variability reduction, the higher the dependency of the feature on that factor. Note that conditioning the feature on an irrelevant factor will not change the distribution. Hence, it has no effect on variability reduction. Figure 3 illustrates the total variability of speech signal as the aggregation of independent factors including lexical, speaker and emotional variabilities (we acknowledge that other aspects not considered here may be relevant). The inner cube represents the total variability space without any restriction on the three factors. This case exhibits the highest level of uncertainty. Conditioning the space with respect to each of the factors will reduce the average variabilities. For instance, speaker-dependent trajectory models for all phonemes and all emotions is expected to have lower variability than a single model for all phonemes, all emotions and all speakers (i.e., the inner cube in Figure 3). These observations motivate our scheme to quantify the dependency between the three given factors and the acoustic features. The mutual information $I(A;F)$ can be used to measure the uncertainty reduction in feature A, given the knowledge on factor F. Mutual information can be expressed in term of the Shannon's entropy $H(A)$ (Equation 1). Notice that $P(f)$ is the probability distribution of factor F, which is estimated from the relative frequencies in the data (e.g., $P(Angry)$ for emotion).

$$I(A;F) = H(A) - \sum_{f \in F} P(f)H(A \mid f) \tag{1}$$

In case of continuous variables, the entropy is replaced by differential entropy. However, differential entropy does not directly determine the variability since *i*) it is scale variant, and *ii*) it can get negative values. Notice that the entropy of multivariate Gaussian distributions reduces to the sum over the logarithms of eigenvalues of the covariance matrix plus a constant term (Cover and Thomas, 2006). This metric is not robust and can cause numerical issues when the covariance matrix has some small eigenvalues, as in this case. In our previous work, a *relevance measure* (RM) was proposed to quantify the uncertainty reduction for continuous multivariate random variables (Equation 2) (Mariooryad and Busso, 2012). In this metric, the uncertainty is measured based on the approach used by Park et al. (2010). The authors used the trace of the covariance matrix (i.e., sum of the eigenvalues) as an approximate uncertainty measure. The eigenvalues of the covariance matrix correspond to the variances observed along the principal directions. Therefore, the sum of eigenvalues, estimated as the trace of the covariance matrix, represents the total variability. $RM(A,F)$ measures the expected uncertainty reduction of feature A when factor F (i.e., speaker identity, lexical content or emotional state) is given. $tr(\Sigma_A)$ is the total variability of feature A, which represents the space covered by the inner cube in Figure 3. For each factor F, we estimate the conditional uncertainty for feature A (i.e., $tr(\Sigma_A|f)$). Then, $RM(A;F)$ gives the expected uncertainty reduction when F is known. We use this metric to quantify the dependency between the factors and the acoustic features (i.e., the higher the value, the stronger the dependency).

$$RM(A;F) = tr(\Sigma_A) - \sum_{f \in F} P(f)tr(\Sigma_A \mid f) \tag{2}$$

5. Factor Analysis Results

5.1. Dominant Factor Across Acoustic Features

The discussed factor analysis technique is conducted on each of the selected features. Notice that the lexical factor is represented at phoneme level. Therefore, phoneme-dependent models, phone level speaker-dependent models and phone level emotion-dependent models are built to calculate the $RM(A;F)$ for these factors. Different features have distinct dynamic ranges and, consequently, different inherent uncertainties. However, the purpose of this analysis is to identify the relative contribution of the three discussed factors on the uncertainty of the acoustic feature A. Therefore, $RM(A;F)$ is normalized across the three factors to sum up to 100. Hence, the reported values quantify the dependency of factor F on feature A in percentage. For each feature, the dominant factor, with the highest dependency, is given in the column "*Factor*". Different colors are used to identify the dominant factor. The lexical, speaker and emotion factors are highlighted with white, light gray and dark gray, respectively.

The result of this analysis is reported in Table 3 (see column "No Normalization"). Some of the coefficients of the *auditory spectrum* (RASTA) and MFCCs present similar patterns (e.g., the second and third MFCCs). Table 3 reports the average results for groups of these coefficients. The last row of the table reports the average RMs across the 57 features considered in the study (i.e., *All*). According to the analysis, the lexical variability across features account for 76% of total variability (on average). The speaker and emotional variabilities account for 14.9% and 9.1% of the total variability, respectively. Due to the

Table 3: Factor analysis results. The table gives the corresponding dependency in percentage, for each feature and factor. The analysis is performed in three conditions: i) without any normalization, ii) with lexical normalization and iii) with speaker normalization. In each case, the dominant factor is highlighted in the Factor column. White, light gray and dark gray represent phoneme-, speaker- and emotion-dependent features, respectively.

	No Normalization				Lexical Normalization				Speaker Normalization			
Feature	Spk	Pho	Emo	Factor	Spk	Pho	Emo	Factor	Spk	Pho	Emo	Factor
auditory spectrum L1 norm	26.06	31.64	42.29	Emotion	39.37	0.87	59.76	Emotion	7.02	42.19	50.78	Emotion
auditory spectrum (RASTA) L1 norm	3.06	87.93	9.01	Phoneme	29.57	13.37	57.07	Emotion	0.72	91.37	7.92	Phoneme
RMS energy	36.75	23.41	39.84	Emotion	48.68	1.23	50.09	Emotion	7.92	37.30	54.78	Emotion
ZCR	0.99	96.86	2.15	Phoneme	33.02	6.02	60.96	Emotion	0.59	97.07	2.33	Phoneme
F0	77.14	1.97	20.89	Speaker	78.27	0.60	21.13	Speaker	13.73	6.40	79.87	Emotion
voicing probability	9.62	81.36	9.01	Phoneme	51.48	3.84	44.68	Speaker	1.99	88.54	9.47	Phoneme
jitter local	11.94	82.31	5.75	Phoneme	30.06	50.82	19.12	Phoneme	2.75	91.91	5.33	Phoneme
jitter delta	22.92	70.75	6.33	Phoneme	0.00	99.24	0.76	Phoneme	5.65	87.93	6.42	Phoneme
auditory spectrum (RASTA) [1-2]	17.12	79.62	3.26	Phoneme	60.62	22.35	17.03	Speaker	1.15	93.88	4.97	Phoneme
auditory spectrum (RASTA) [3-10]	2.41	93.00	4.57	Phoneme	26.70	19.64	53.66	Emotion	0.58	95.41	4.01	Phoneme
auditory spectrum (RASTA) [11-18]	2.19	91.50	6.31	Phoneme	25.79	18.76	55.45	Emotion	0.57	93.72	5.72	Phoneme
auditory spectrum (RASTA) [19-26]	2.34	93.42	4.24	Phoneme	30.23	21.41	48.35	Emotion	0.64	95.20	4.17	Phoneme
MFCC [1]	4.21	90.95	4.84	Phoneme	45.97	6.28	47.75	Emotion	0.98	94.24	4.78	Phoneme
MFCC [2-8]	23.37	69.76	6.87	Phoneme	73.07	3.47	23.46	Speaker	3.90	87.71	8.38	Phoneme
MFCC [9-12]	57.46	35.32	7.22	Speaker	83.89	4.50	11.61	Speaker	12.25	80.31	7.43	Phoneme
spectral components fband 25-650	49.19	13.83	36.99	Speaker	56.69	2.83	40.49	Speaker	11.26	28.28	60.46	Emotion
spectral components fband 1000-4000	43.18	17.58	39.24	Speaker	54.53	5.34	40.13	Speaker	6.35	39.94	53.71	Emotion
spectral components roll off 25%	13.52	65.75	20.73	Phoneme	49.08	1.52	49.40	Emotion	3.73	69.34	26.93	Phoneme
spectral components roll off 50%	4.33	85.85	9.82	Phoneme	43.19	2.58	54.23	Emotion	2.02	86.20	11.78	Phoneme
spectral components roll off 75%	2.21	91.68	6.11	Phoneme	38.38	3.46	58.17	Emotion	1.57	90.64	7.79	Phoneme
spectral components roll off 90%	1.72	93.30	4.98	Phoneme	36.63	3.71	59.66	Emotion	1.45	92.17	6.38	Phoneme
spectral components flux	34.89	16.11	48.99	Emotion	42.42	1.25	56.33	Emotion	8.22	26.24	65.54	Emotion
spectral components entropy	11.57	87.95	0.49	Phoneme	91.25	5.46	3.29	Speaker	1.31	98.30	0.39	Phoneme
spectral components variance	0.72	98.82	0.46	Phoneme	42.71	15.32	41.96	Speaker	0.33	99.27	0.40	Phoneme
spectral components skewness	4.13	92.42	3.46	Phoneme	49.36	11.74	38.90	Speaker	1.65	94.77	3.57	Phoneme
spectral components kurtosis	7.66	89.52	2.82	Phoneme	66.36	8.63	25.01	Speaker	3.21	93.84	2.94	Phoneme
Average [*Selected*]	9.53	79.47	11.00	Phoneme	33.99	14.26	51.75	Emotion	2.16	83.13	14.71	Phoneme
Average [*All*]	14.89	76.01	9.10	Phoneme	44.86	14.20	40.94	Speaker	3.08	85.15	11.77	Phoneme

segmental nature of the analysis, the relevance measure captures the close relationship between the phonetic content and the acoustic features. If the analysis had considered longer, suprasegmental lexical units (e.g., phrases), these percentages could have been different. Table 3 shows that the F0 contour is highly dependent on the speaker factor (77.1%). This result is expected given the differences observed in the vocal fold structure across individuals. This table shows that all the auditory spectrum components and the first eight MFCCs are mainly dependent on the lexical content. These speech features are the most informative and often used features for *automatic speech recognition* (ASR). The 9^{th}-12^{th} MFCCs and the energy in low (*fband 25-650Hz*) and high (*fband 1-4Hz*) bands are highly dependent on the speaker. These features are good candidates for speaker identification tasks. According to this table, emotion is the dominant factor in some energy related features, including *RMS energy*, *auditory spectrum L1 norm* and *spectral flux*. The variabilities of the remaining features are all dominated by the lexical content. *ZCR* is one of these lexical-dependent features. This result is also expected, given the discriminative power of *ZCR* to distinguish between voiced and unvoiced speech (Rabiner and Sambur, 1975).

5.2. Speaker and Lexical Normalization

According to Table 3, emotion is not the dominant factor for most of the acoustic features. However, it modulates the trajectory shape of certain features. For instance, 20.9% of the F0 variability can be attributed to the emotional content (as measured by the proposed RM). Similarly, the emotion factor introduces a variability on features that are mostly dependent on the lexical content (e.g., *auditory spectrum (RASTA) L1 norm* and *ZCR*). These results suggest that attenuating the speaker- and lexical-dependent variations can enhance the characterization of emotional information. To address this hypothesis, we propose a normalization scheme to remove the speaker and lexical variabilities. The proposed normalization is based on our previous work on lexical normalization for facial expressions (Mariooryad and Busso, 2013b). It uses the whitening transformation over the trajectory models introduced in Section 4.1. While this section discusses the lexical normalization by employing the phoneme-dependent models, the same arguments and derivations are applicable for speaker normalization with the speaker-dependent models. Given the phoneme-dependent trajectory models, a phoneme is selected as reference $p_{ref} \sim N(\mu_{ref}, \Sigma_{ref})$. Then, the goal is to transform each phoneme $p_i \sim N(\mu_i, \Sigma_i)$ to have the same statistics as the reference phoneme. Given $X_{N\times 1}$, a sample of p_i, equations 3 and 4 perform this transformation. V_i and D_i are the eigenvectors and eigenvalues of Σ_i in matrix form. Similarly, V_{ref} and D_{ref} are the eigenvectors and eigenvalues of Σ_{ref}. The whitening transformation in equation 3 decorrelates the elements of X. This equation is the extension of the conventional Z normalization for multivariate random variables. After applying the whitening transformation, the next step is to impose the statistics of p_{ref} on the decorrelated data

with equation 4 (i.e., coloring transformation). Then, all the phonemes will have similar first and second order statistics as the reference phoneme. Therefore, the lexical variability is attenuated by these linear transformations. Motivated by our previous works (Busso et al., 2007, 2009; Mariooryad and Busso, 2013b) the transformation parameters are estimated from the neutral utterances of the database. Hence, after performing the whitening-coloring steps, all the neutral phonemes are forced to have similar statistics, regardless of the actual phonetic label. Deviation from this unified neutral distribution will signal the existence of emotional behaviors.

$$X^w = D_i^{-\frac{1}{2}} V_i'(X - \mu_i) \tag{3}$$

$$X^n = V_{ref} D_{ref}^{\frac{1}{2}} X^w + \mu_{ref} \tag{4}$$

The first and second rows of Figure 4 show the trajectory models of *auditory spectrum L1 norm* for phonemes /æ/ and /ə/, respectively. The third row of Figure 4 depicts the trajectories of phoneme /æ/, after the whitening-coloring normalization (i.e., /æ/ ⇒ /ə/). The phoneme /ə/ is considered as the reference phoneme in this normalization. Notice that the normalized trajectory for neutral emotion is exactly the same as the neutral trajectory of phoneme /ə/ (see Fig. 4(e) and Fig. 4(i)). Likewise, considering the corresponding emotions, the similarity between trajectories of /ə/ and /æ/ ⇒ /ə/ increases compared to the pair /ə/ and /æ/. Notice that the trajectory of /æ/ for happiness (Fig. 4(b)) is more similar to /ə/ in anger (Fig. 4(g)) than to /ə/ for happiness (Fig. 4(f)). These figures show the challenge that the underlying lexical content can introduce, blurring the emotion information. However, the proposed normalization reduces this nuisance variability. Notice that after the normalization the trajectory of /æ/ ⇒ /ə/ in happiness is best matched with the trajectory of /ə/ in happiness (Fig. 4(j) and Fig. 4(f)). This pattern is also consistent across other emotions (compare the second and third rows of Fig. 4).

The column "*Lexical Normalization*" in Table 3 gives the results of the factor analysis method after lexical normalization. Since /ə/ is the most frequent phoneme in the IEMOCAP, it is chosen as the reference phoneme for lexical normalization. The last row of Table 3 shows that after the normalization the average dependency on lexical content, measured by the proposed RM, reduces from 76% to 14.2%. Hence, the variabilities of speaker and emotional factors increase to 44.9% and 40.9%, respectively. These results show the effectiveness of the proposed whitening normalization. After lexical normalization, the lexical factor is not selected as the dominant factor in any of the features. After lexical normalization, the features ZCR, the first MFCC and most of the auditory spectral components are mainly dependent on the emotion factor. Speaker factor becomes the dominant factor in the rest of the MFCCs. Although some studies have used MFCCs for emotion recognition (Metallinou et al., 2012), this result suggests that MFBs are more discriminative for emotions. The low pass liftering and the decorrelation by the *discrete cosine transformation* (DCT) seems to affect the discriminative power of the features for characterizing emotions. Our previous experiments also support this finding (Busso et al., 2007).

A similar normalization scheme is adapted to compensate for the speaker factor. The speaker-dependent models across all phonemes are used to perform the whitening-coloring scheme to remove the speaker dependency. Notice that due to limited size of the corpus, it is not practical to simultaneously perform speaker-dependent and phoneme-dependent normalizations. Therefore, these two normalizations are separately studied. The column "Speaker Normalization" in Table 3 gives the results of the factor analysis method after applying the speaker normalization. The female subject in the first session of the database is selected as the reference speaker. According to the last row of this column, the proposed speaker normalization reduces the speaker dependency, across all features, from 14.9% to 3.1%, increasing the dependency on lexical and emotional factors. On average, the trajectory-based speaker normalization approach improves the emotion RM from 9.1% to 11.8%. Notice that it is not as effective as the lexical normalization (i.e., from 9.1% to 40.9%). However, it still increases the dependency on the emotion factor. For instance, it makes emotion the dominant factor on F0, energy in low frequency (i.e, fband 25-650Hz) and high frequency (i.e., fband 1k-4kHz) bands. F0 is one of the most commonly used features in emotion recognition studies. This result not only supports the benefit of using this feature, but also validates the expectation that speaker normalization can improve emotion discrimination.

In Table 3, the features yielding emotion as the dominant factor before or after any of the normalizations are highlighted in dark gray. This set comprises of 37 features, which is referred to as the *Selected* feature set. Notice that in the *Selected* set, the lexical normalization yields higher RM values across features for the emotional factor than the ones for the speaker factor. Therefore, it is expected that these features are more relevant for emotion recognition.

6. Emotion Classification

This section leverages the insights from the factor analysis to improve the robustness of an emotion recognition system. The experiments are performed in 10 fold leave-one-speaker-out cross validation scheme. Each fold has two speaker-independent partitions (9 speakers for training and 1 speaker for testing). Emotion recognition experiments demonstrate the discriminative power of the features selected by the analysis, along with the effectiveness of the lexical and speaker normalizations.

The *Selected* feature set presented in Section 5.2 is obtained using all the database including training and testing partitions. For the classification experiments, the feature selection should be conducted only over the training partition. Therefore, we repeated the analysis per fold using only the training partition (the testing partition is not considered). This approach produces ten feature sets. On average, 35.5 (σ = 1.5) low level descriptors are selected for each fold. 32 features are consistently selected across the folds. In the rest of this paper, the *Selected* feature set refers to these feature sets. Likewise, we train the

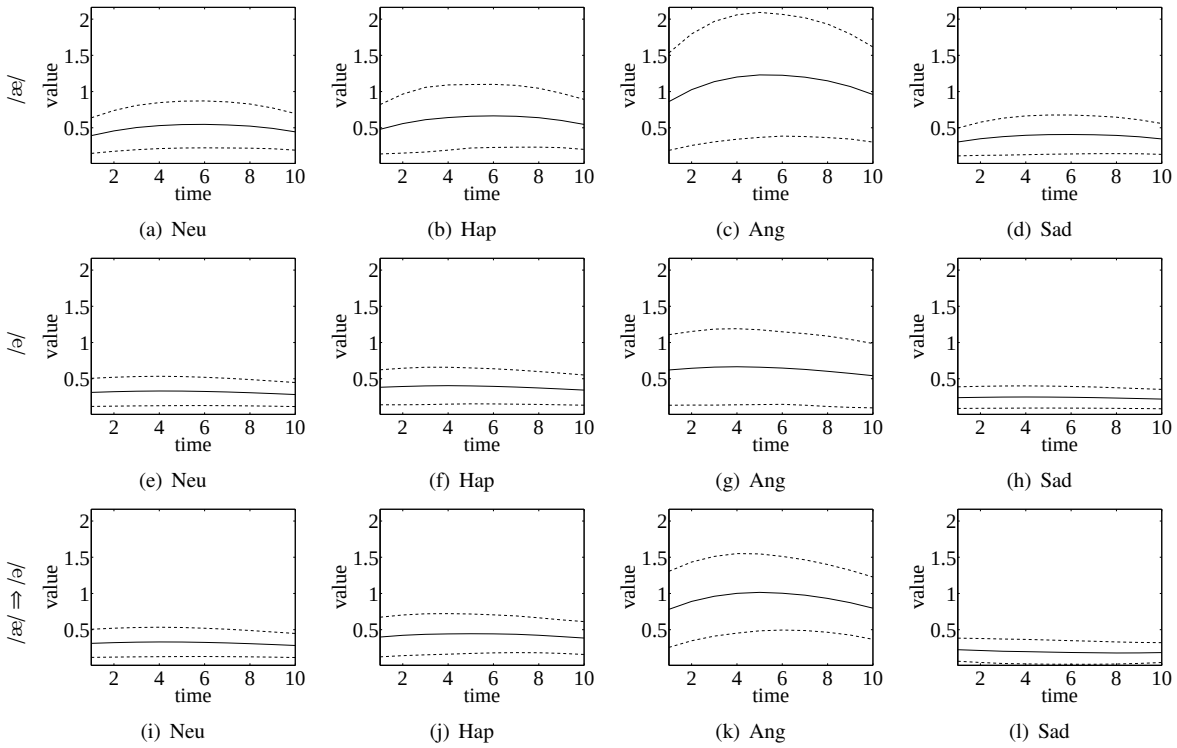

Figure 4: Mean and standard deviation of the trajectories of the *auditory spectrum L1 norm*. The first and second rows depict the trajectory models of phonemes /æ/ and /ə/, respectively. The third row shows the trajectories of /æ/ after normalization, with respect to the reference phoneme /ə/. For each emotion, the trajectory is extracted across all speakers. Neu: neutral, Hap: happiness, Ang: anger, Sad: sadness

classifiers using all the features (set *All*). For each feature, a set of nine high level statistics are extracted representing the utterance. The set of statistics include: minimum, maximum, mean, median, standard deviation, kurtosis, skewness, lower quartile and upper quartile. These statistics were carefully selected to minimize the effect of the discontinuities introduced at the phoneme boundaries by the normalization schemes (the phonemes are independently normalized using Equation 3 – see Sec. 6.2).

We use *support vector machine* (SVM) with linear kernel to perform the classifications. The WEKA data mining package (Hall et al., 2009) is used to train the SVMs with *sequential minimal optimization* (SMO). The *complexity* parameter of the classifier c is set to 1 for all the settings. Our preliminary experiments showed that lower values of c degrade the emotion recognition performance in all the settings. Due to the unbalanced emotional classes (see Table 1), the results are presented in terms of *accuracy* (A), average *precision* (P), average *recall* (R) and *F-score* (F). Precision of an emotional class is the fraction of relevant samples that are correctly classified. We estimate the average precision, achieved by each emotional class. Recall is the fraction of correctly retrieved samples for one emotional class. We estimate the average recall across the emotional classes. Equation 5 gives the *F-score*, which combines both metrics.

$$F = \frac{2PR}{P+R} \tag{5}$$

6.1. Baseline

Table 4 gives the baseline performance achieved by training the classifier with acoustic features without any normalization. A previous study on a smaller portion of this database reported 50.69% of average recall for this task, using an HMM-based method (Metallinou et al., 2010b). Li et al. (2012) have reported 54.34% and 55.84% average recall with GMM-based approach and latent factor analysis technique, respectively. The low accuracies are explained by the ambiguous emotional content of this spontaneous corpus (Mower et al., 2009). For example, only 41% of the samples have a complete consensus in the labels assigned by the raters in the subjective evaluations conducted over the portion of the database used in our study. In spite of the inherent ambiguity of the task, our SVM baseline provides a competitive performance compared to previous studies on this corpus.

The experiments are performed both on the *Selected* and *All* feature sets. Interestingly, the differences in performance of the classifiers trained with *All* and *Selected* feature sets are not statistically significant in terms of any of the performance metrics ($p - value > 0.082$). This result suggests that the proposed shape-based analysis can serve as a principled method for feature selection. This is an important problem in automatic emotion recognition, since the classifiers are commonly trained with high dimensional feature vectors obtained by extracting multiple statistics over the low level descriptors. This approach can identify low level features that are not affected by emotions, reducing the high level descriptors measurably. For example, the

Table 4: Baseline emotion recognition performance. The results are reported in terms of *Accuracy* (A) and average *precision* (P), average *recall* (R) and F-score (F).

Feature Set	A	P	R	F
Selected	54.05	54.34	54.23	54.28
All	55.32	55.95	55.73	55.84

Table 5: Emotion recognition performance after the lexical normalization. The results are reported in terms of *Accuracy* (A) and average *precision* (P), average *recall* (R) and F-score (F).

Feature Set	A	P	R	F
Selected	56.24	57.10	56.13	56.61
All	56.75	57.47	56.73	57.10

classifiers built with the set *All* contain 513 (57 acoustic features $\times$ 9 statistics). In contrast, the classifiers trained with the *Selected* set contain, on average, 319.5 (35.5 acoustic features $\times$ 9 statistics) features. This corresponds to 38% feature reduction.

6.2. Effect of Lexical Normalization

To evaluate the effect of lexical normalization, the proposed whitening transformation scheme is applied on each of the phonemes in a given utterance (i.e., the whitening step). A single garbage model is built for the phonemes with less than 100 repetitions per emotion. The shape based lexical normalization is implemented as follow. First, each phoneme is independently normalized using the proposed factor analysis (Sec. 5.2). Here, we assume that the phonetic transcription is known. Then, the normalized trajectories are concatenated, preserving the time duration (i.e., the whitened phonemes are resampled to preserve the initial phoneme durations). Finally, we estimated the same high level statistics used for the baseline classifiers at the sentence level. These statistics are used as feature in the experiments. The proposed normalization uses a different linear transformation for each phoneme (i.e., phoneme-dependent). Hence, there are no discontinuities within a phoneme. However, by applying different linear transformations on adjacent phonemes, the approach introduces discontinuities at the phoneme boundaries on the features (i.e., the smooth transition between the phoneme boundaries is not necessarily preserved after the normalization step). Notice that these discontinuities do not affect the features since the statistics are carefully selected to avoid incorporating the effect of breaks. We do not consider statistics derived from features' derivative and common functionals used in other studies such as slope or linear regression coefficients.

Table 5 reports the recognition performance after the lexical normalization. With the *Selected* feature set, the effect of lexical normalization is statistically significant for A, P and F ($p - value < 0.03$). The results correspond to 4.1% and 4.3% relative improvement over the baseline for A and F, respectively. These results support the effectiveness of the proposed whitening transformation to compensate for the lexical variability. Note that this normalization also improves the performance on the classifier trained with all features.

Table 6: Emotion recognition performance after the speaker normalization. The results are reported in terms of *Accuracy* (A) and average *precision* (P), average *recall* (R) and F-score (F).

Feature Set	A	P	R	F
Selected	55.37	56.34	55.37	55.85
All	54.45	55.27	54.67	54.97

6.3. Effect of Speaker Normalization

The discussed experimental settings are adapted to analyze the effect of speaker normalization with the whitening transformation for emotion recognition. Hence, for each speaker, the corresponding speaker-dependent model is used to whiten the phonemes in the utterance. Table 6 gives the results of this experiment. Notice that with the *Selected* feature set, the performance increases. The result gives a 2.4% relative improvement in accuracy compared to the baseline. However, using the feature set *All* reduces the recognition rates. According to Table 3, speaker normalization boosts both lexical and emotional variabilities. However, the lexical variability is still the dominant factor across the features. In fact, on average, the emotion RM only increases from 9.1% to 11.8%. This effect is more evident in the feature set *All*. Hence, it is expected to see a drop in performance when all the features are used. We conclude that the proposed trajectory-based normalization is more suitable to compensate for the lexical information.

7. Conclusions and Discussions

This paper introduced a factor analysis framework to identify the dependency of various acoustic features on speaker-dependent characteristics, lexical content and emotional states. For this purpose, a trajectory model was developed to analyze the variability in the temporal shape of an exhaustive set of acoustic features at the phoneme level. The factor analysis metric quantifies the dependency of each feature on the speaker, lexical, and emotion aspects by measuring the reduction in uncertainty associated to the knowledge of a given factor. On average, the analysis showed that 76% of the variability in the trajectories was associated with the lexical content. Motivated by the analysis, we introduced a lexical/speaker normalization scheme based on the whitening transformation. After the lexical/speaker normalization, the variability associated to the normalized factor significantly decreases, showing the effectiveness of the technique to compensate for the speaker and lexical variabilities.

The analysis and the normalization scheme were validated with emotion classification experiments. We built two baseline classifiers trained with either all the features or a reduced features set. This selected features set included all the features for which the emotional factor was the primary source of variability, before or after the speaker or lexical normalization. The proportion hypothesis test shows no significant difference in the performance of both classifiers, even when 38% less features were used with the *Selected* set. The results suggest that the proposed shape-based analysis can be used as a

principled method to eliminate low level descriptors that do not provide emotional information. Likewise, emotion classification experiments show that the proposed lexical normalization scheme produces a relative improvement of 4.1% over the baseline, when the *Selected* feature set is used. The best average recall achieved in this study is 56.73%, which is higher than the performance reported in previous studies on this corpus (Metallinou et al., 2010b; Li et al., 2012).

The analysis in this paper provides a systematic way to identify acoustic features in which the variability in their trajectories are mainly associated with emotional information (before or after the speaker/lexical normalization). One interesting direction is to set more restrictive thresholds on the relevance metrics to reduce even more the number of features. Notice that the feature selection is done at the low level descriptors, before estimating the statistics, which is a novel aspect of the approach. Likewise, we are working on identifying other machine learning framework that may better leverage the proposed feature normalization (e.g., continuous frame-by-frame classifiers).

When the proposed normalization scheme is used to compensate for the speaker variability, the emotion recognition results do not provide significant improvement. This result is not surprising, since the percentage of variability explained by the emotional factor only increases from 9.1% to 11.8% after the speaker normalization. Given that previous studies have shown the importance of speaker normalization (Busso et al., 2011), we conclude that shape-based, phoneme-level normalization may not be the best approach to reduce the speaker variability. We are currently exploring alternative lexical-independent approaches based on global statistics to normalize the speaker differences that can be coupled with the proposed lexical normalization scheme. We will consider linear and nonlinear normalization methods (Sethu et al., 2007).

The insights of the analysis can be leveraged in various speech processing tasks. For example, the normalization approach can be used to compensate for the emotion variability in the context of automatic speech recognition and speaker identification. Studies have shown that the performance of these systems significantly drops in presence of emotions (Wu et al., 2006; Schuller et al., 2006) and different speaking styles (Shriberg et al., 2008). Therefore, a feature normalization scheme can be used to reduce the emotional variability, overcoming these problems.

Since this study assumes that the underlying lexical content is given, the set of potential applications is limited (e.g., judicial recordings with available transcriptions). In other applications, the proposed method should rely on transcripts generated by ASR systems. We are planning to analyze the performance of the proposed method in presence of recognition errors introduced by ASR systems. Alternatively, we are planning to study factor analysis models such as bilinear model (Tenenbaum and Freeman, 2000) and tied factor analysis based on *probabilistic linear discriminant analysis* (PLDA) (Prince et al., 2008) to tackle this issue. With these factor separation models, the phoneme labels are only required during the training, to model the relationship between lexical content, emotions and the acoustic features. The testing step utilizes the trained models and considers the lexical content as a missing variable to infer the underlying emotion, without given transcripts. These research directions are important steps toward building robust emotion recognition systems that can be successfully deployed in practical applications.

Acknowledgment

This study was funded by Samsung Telecommunications America and the US National Science Foundation under grants IIS 1217104 and IIS: 1329659.

References

Batliner, A., Seppi, D., Steidl, S., Schuller, B., 2010. Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach. Advances in Human-Computer Interaction 2010, 1–15.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S., Narayanan, S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation 42, 335–359.

Busso, C., Lee, S., Narayanan, S., 2007. Using neutral speech models for emotional speech analysis, in: Interspeech 2007 - Eurospeech, Antwerp, Belgium. pp. 2225–2228.

Busso, C., Lee, S., Narayanan, S., 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Transactions on Audio, Speech and Language Processing 17, 582–596.

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

Busso, C., Narayanan, S., 2006. Interplay between linguistic and affective goals in facial expression during emotional utterances, in: 7th International Seminar on Speech Production (ISSP 2006), Ubatuba-SP, Brazil. pp. 549–556.

Busso, C., Narayanan, S., 2007. Joint analysis of the emotional fingerprint in the face and speech: A single subject study, in: International Workshop on Multimedia Signal Processing (MMSP 2007), Chania, Crete, Greece. pp. 43–47.

Chappell, D., Hansen, J., 1998. Speaker-specific pitch contour modeling and modification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998), Seattle, WA, USA. pp. 885–888.

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

Cover, T., Thomas, J., 2006. Elements of Information Theory. Wiley-Interscience, New York, NY, USA.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., Dumouchel, P., 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, in: Interspeech 2009, Brighton, UK. pp. 1559–1562.

Dehak, N., Torres-Carrasquillo, P., Reynolds, D., Dehak, R., 2011. Language recognition via i-vectors and dimensionality reduction, in: Interspeech 2011, Florence, Italy. pp. 857–860.

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

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

Gish, H., Ng, K., 1993. A segmental speech model with applications to word spotting, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1993), Minneapolis, MN, USA. pp. 447–450.

Gish, H., Ng, K., 1996. Parametric trajectory models for speech recognition, in: Proceedings of the Fourth International Conference on Spoken Language (ICSLP 1996), Philadelphia, PA, USA. pp. 466–469.

Gong, Y., 1997. Stochastic trajectory modeling and sentence searching for continuous speech recognition. IEEE Transactions on Speech and Audio Processing 5, 33–44.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., 2009. The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter 11, 10–18.

Hansen, J., Womack, B., 1996. Feature analysis and neural network-based classification of speech under stress. IEEE Transactions on Speech and Audio Processing 4, 307–313.

Hastie, R., Dawes, R.M., 2001. Rational choice in an uncertain world: The Psychology of Judgement and Decision Making. Sage Publications, Inc, Thousand Oaks, CA, USA.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America 87, 1738–1752.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 15, 1448–1460.

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

Lee, S., Yildirim, S., Kazemzadeh, A., Narayanan, S., 2005. An articulatory study of emotional speech production, in: 9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech), Lisbon, Portugal. pp. 497–500.

Lei, X., Siu, M., Hwang, M., Ostendorf, M., Lee, T., 2006. Improved tone modeling for mandarin broadcast news speech recognition, in: International Conference on Spoken Language (ICSLP 2006), Pittsburgh, PA, USA. pp. 1237–1240.

Li, M., Metallinou, A., Bone, D., Narayanan, S., 2012. Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan. pp. 1937–1940.

Mariooryad, S., Busso, C., 2012. Factorizing speaker, lexical and emotional variabilities observed in facial expressions, in: IEEE International Conference on Image Processing (ICIP 2012), Orlando, FL, USA. pp. 2605–2608.

Mariooryad, S., Busso, C., 2013a. Exploring cross-modality affective reactions for audiovisual emotion recognition. IEEE Transactions on Affective Computing In Press.

Mariooryad, S., Busso, C., 2013b. Feature and model level compensation of lexical content for facial emotion recognition, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013), Shanghai, China.

Metallinou, A., Busso, C., Lee, S., Narayanan, S., 2010a. Visual emotion recognition using compact facial representations and viseme information, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Dallas, TX, USA. pp. 2474–2477.

Metallinou, A., Katsamanis, A., Narayanan, S., 2012. 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. pp. 2401–2404.

Metallinou, A., Lee, S., Narayanan, S., 2010b. Decision level combination of multiple modalities for recognition and analysis of emotional expression, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Dallas, TX, USA. pp. 2462–2465.

Mower, E., Metallinou, A., Lee, C.C., Kazemzadeh, A., Busso, C., Lee, S., Narayanan, S., 2009. Interpreting ambiguous emotional expressions, in: International Conference on Affective Computing and Intelligent Interaction (ACII 2009), Amsterdam, The Netherlands.

Park, S., Choi, H., Roy, N., How, J., 2010. Learning the covariance dynamics of a large-scale environment for informative path planning of unmanned aerial vehicle sensors. International Journal of Aeronautical and Space Sciences 11, 327–337.

Prince, S., Elder, J., Warrell, J., Felisberti, F., 2008. Tied factor analysis for face recognition across large pose differences. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 970–984.

Rabiner, L., Sambur, M., 1975. An algorithm for detecting the endpoints of isolated utterances. Bell System Technical Journal 54, 297–315.

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

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011a. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53, 1062–1087.

Schuller, B., Stadermann, J., Rigoll, G., 2006. Affect-robust speech recognition by dynamic emotional adaptation, in: ISCA Speech Prosody 2006, ISCA, Dresden, Germany.

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

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G., 2010. Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Transactions on Affective Computing 1, 119–131.

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

Shriberg, E., Graciarena, M., Bratt, H., Kathol, A., Kajarekar, S., Jameel, H., Richey, C., Goodman, F., 2008. Effects of vocal effort and speaking style on text-independent speaker verification, in: Interspeech 2008, Brisbane, Australia. pp. 609–612.

Tenenbaum, J., Freeman, W., 2000. Separating style and content with bilinear models. Journal of Neural Computation 12, 1247–1283.

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

Vlasenko, B., Prylipko, D., Philippou-Hübner, D., Wendemuth, A., 2011b. 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. pp. 1577–1580.

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

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

Wu, W., Zheng, T., Xu, M., Bao, H., 2006. Study on speaker verification on emotional speech, in: International Conference on Spoken Language (ICSLP 2006), Pittsburgh, PA, USA. pp. 2102–2105.

Xia, R., Liu, Y., 2012. Using i-vector space model for emotion recognition, in: Interspeech 2012, Portland, Oregon, USA. pp. 2230–2233.

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

Zeng, Z., Tu, J., Pianfetti, B., Huang, T., 2008. Audiovisual affective expression recognition through multistream fused HMM. IEEE Transactions on Multimedia 10, 570–577.