ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication xxx (2010) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic speech emotion recognition using modulation spectral features

Siqing Wu [a,*], Tiago H. Falk [b], Wai-Yip Chan [a]

[a] Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada K7L 3N6
[b] Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada M5S 3G9

Abstract

In this study, modulation spectral features (MSFs) are proposed for the automatic recognition of human affective information from speech. The features are extracted from an auditory-inspired long-term spectro-temporal representation. Obtained using an auditory filterbank and a modulation filterbank for speech analysis, the representation captures both acoustic frequency and temporal modulation frequency components, thereby conveying information that is important for human speech perception but missing from conventional short-term spectral features. On an experiment assessing classification of discrete emotion categories, the MSFs show promising performance in comparison with features that are based on mel-frequency cepstral coefficients and perceptual linear prediction coefficients, two commonly used short-term spectral representations. The MSFs further render a substantial improvement in recognition performance when used to augment prosodic features, which have been extensively used for emotion recognition. Using both types of features, an overall recognition rate of 91.6% is obtained for classifying seven emotion categories. Moreover, in an experiment assessing recognition of continuous emotions, the proposed features in combination with prosodic features attain estimation performance comparable to human evaluation.

Keywords: Emotion recognition; Speech modulation; Spectro-temporal representation; Affective computing; Speech analysis

1. Introduction

Affective computing, an active interdisciplinary research field, is concerned with the automatic recognition, interpretation, and synthesis of human emotions (Picard, 1997). Within its areas of interest, speech emotion recognition (SER) aims at recognizing the underlying emotional state of a speaker from the speech signal. The paralinguistic information conveyed by speech emotions has been found to be useful in multiple ways in speech processing, especially serving as an important ingredient of "emotional intelligence" of machines and contributing to human–machine interaction (Cowie et al., 2001; Ververidis and Kotropoulos, 2006). Moreover, since a broad range of emotions can be faithfully delivered in a telephone conversation where only auditory information is exchanged, it should be possible to build high-performance emotion recognition systems, using only speech signals as the input. Such speech based systems can function either independently or as modules of more sophisticated techniques that combine other information sources such as facial expression and gesture (Gunes and Piccard, 2007).

Despite the substantial advances made in this area, SER still faces a number of challenges, one of which is designing effective features. Most acoustic features that have been used for emotion recognition can be divided into two categories: prosodic and spectral. Prosodic features have been shown to deliver important emotional cues of the speaker (Cowie et al., 2001; Ververidis and Kotropoulos, 2006; Busso et al., 2009). Even though there is no agreement on the best features to use, prosodic features form the most

* Corresponding author.
E-mail addresses: siqing.wu@queensu.ca (S. Wu), chan@queensu.ca (W.-Y. Chan).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.013

commonly used feature type for SER, and have been extensively studied by previous works (e.g. Cowie and Douglas-Cowie, 1996; Abelin and Allwood, 2000; Cowie et al., 2001; Mozziconacci, 2002; Scherer, 2003; Barra et al., 2006; Schuller et al., 2007a; Busso et al., 2009). On the other hand, spectral features (including cepstral features) also play a significant role in SER as they convey the frequency content of the speech signal, and provide complementary information to prosodic features. Comparatively, however, limited research efforts have been put into constructing more powerful spectral features for emotion recognition. The spectral features are usually extracted over a short frame duration (e.g. 20–30 ms), with longer temporal information incorporated in the form of local derivatives (e.g. Nwe et al., 2003; Batliner et al., 2006; Vlasenko et al., 2007).

The limitations of short-term spectral features for speech recognition, however, are considerable (Morgan et al., 2005). Even with the inclusion of local derivatives, the fundamental character of the features remains fairly short-term. In short, conventional spectral features used for speech recognition, such as the well-known mel-frequency cepstral coefficients (MFCCs), convey the signal's short-term spectral properties only, omitting important temporal behavior information. Such limitations are also likely to hamper SER performance. On the other hand, advances in neuroscience suggest the existence of spectro-temporal (ST) receptive fields in mammalian auditory cortex which can extend up to temporal spans of hundreds of milliseconds and respond to modulations in the time-frequency domain (Depireux et al., 2001; Shamma, 2001; Chih et al., 2005). The importance of the modulation spectrum of speech is evident in a number of areas, including auditory physiology, psychoacoustics, speech perception, and signal analysis and synthesis, as summarized in (Atlas and Shamma, 2003). These new insights further reveal the shortcomings of short-term spectral features as they discard the long-term temporal cues used by human listeners, and highlight the need for more perceptually motivated features.

In line with these findings, long-term modulation spectral features (MSFs) are proposed in this paper for emotion recognition. These features are based on frequency analysis of the temporal envelopes (amplitude modulations) of multiple acoustic frequency bins, thus capturing both spectral and temporal properties of the speech signal. The proposed features are applied to two different SER tasks: (1) classification of *discrete emotions* (e.g. *joy*, *neutral*) under the categorical framework which characterizes speech emotions using categorical descriptors and (2) estimation of *continuous emotions* (e.g. *valence*, *activation*) under the dimensional framework which describes speech emotions as points in an emotion space. In the past, classification tasks have drawn dominant attention of the research community (Cowie et al., 2001; Douglas-Cowie et al., 2003; Ververidis and Kotropoulos, 2006; Shami and Verhelst, 2007). Recent studies, however, have also focused on recognizing continuous emotions (Grimm et al., 2007a,b; Wollmer et al., 2008; Giannakopoulos et al., 2009).

To our knowledge, the only previous attempt at using modulation spectral content for the purpose of emotion recognition is reported in (Scherer et al., 2007), where the modulation features are combined with several other feature types (e.g., loudness features) and approximately 70% recognition rate is achieved on the so-called Berlin emotional speech database (Burkhardt et al., 2005). This present study, which extends our previous work (Wu et al., 2009), is different in several ways, namely (1) filterbanks are employed for spectral decomposition; (2) the proposed MSFs are designed by exploiting a long-term ST representation of speech, and are shown to achieve considerably better performance on the Berlin database relative to (Scherer et al., 2007); and (3) continuous emotion estimation is also performed.

The remainder of the paper is organized as follows. Section 2 presents the algorithm for generating the long-term ST representation of speech. Section 3 details the MSFs proposed in this work, as well as short-term spectral features and prosodic features extracted for comparison purposes. Section 4 introduces the databases employed. Experimental results are presented and discussed in Section 5, where both discrete emotion classification (Section 5.1) and continuous emotion estimation (Section 5.2) are performed. Finally, Section 6 gives concluding remarks.

2. ST representation of speech

The auditory-inspired spectro-temporal (ST) representation of speech is obtained via the steps depicted in Fig. 1. The initial pre-processing module resamples the speech signal to 8 kHz and normalizes its active speech level to −26 dBov using the P.56 speech voltmeter (Intl. Telecom. Union, 1993). Since emotions can be reliably conveyed through band-limited telephone speech, we consider the 8 kHz sampling rate adequate for SER. Speech frames (without overlap) are labeled as active or inactive by the G.729 voice activity detection (VAD) algorithm described in (Intl. Telecom. Union, 1996) and only active speech frames are retained. The preprocessed speech signal $s(n)$ is framed into long-term segments $s_k(n)$ by multiplying a 256 ms Hamming window with 64 ms frame shift, where k denotes the frame index. Because the first subband filter in the modulation filterbank (described below) analyzes frequency content around 4 Hz, this relatively long temporal span is necessary for such low modulation frequencies.

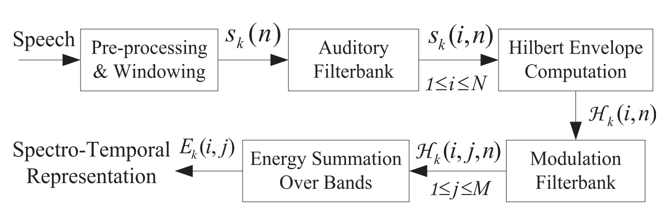

Fig. 1. Flowchart for deriving the ST representation.

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

It is well-known that the human auditory system can be modeled as a series of over-lapping band-pass frequency channels (Fletcher, 1940), namely auditory filters with critical bandwidths that increase with filter center frequencies. The output signal of the ith critical-band filter at frame k is given by:

$$s_k(i,n) = s_k(n) * h(i,n), \tag{1}$$

where $h(i,n)$ denotes the impulse response of the ith channel, and * denotes convolution. Here, a critical-band gammatone filterbank (Aertsen and Johannesma, 1980) with N subband filters is employed. The implementation in (Slaney, 1993) is used. The center frequencies of these filters (namely *acoustic* frequency, to distinguish from *modulation* frequency of the modulation filterbank) are proportional to their bandwidths, which in turn, are characterized by the equivalent rectangular bandwidth (Glasberg and Moore, 1990):

$$ERB_i = \frac{F_i}{Q_{ear}} + B_{min}, \tag{2}$$

where F_i is the center frequency (in Hz) of the ith critical-band filter, and Q_{ear} and B_{min} are constants set to 9.26449 and 24.7, respectively. In our simulations, a gammatone filterbank with 19 filters is used, where the first and the last filters are centered at 125 Hz and 3.5 kHz, with bandwidths of 38 and 400 Hz, respectively. The magnitude response of the filterbank is depicted in Fig. 2.The temporal envelope, or more specifically, the Hilbert envelope $\mathcal{H}_k(i,n)$, is then computed from $s_k(i,n)$ as the magnitude of the complex analytic signal $\hat{s}_k(i,n) = s_k(i,n) + j\mathbb{H}\{s_k(i,n)\}$, where $\mathbb{H}\{\cdot\}$ denotes the Hilbert transform. Hence,

$$\mathcal{H}_k(i,n) = |\hat{s}_k(i,n)| = \sqrt{s_k^2(i,n) + \mathbb{H}^2\{s_k(i,n)\}}. \tag{3}$$

Fig. 3 shows an example of a bandpassed speech segment (subplot a) and its Hilbert envelope (subplot b).

The auditory spectral decomposition modeled by the critical-band filterbank, however, only comprises the first stage of the signal transformation performed in the human auditory system. The output of this early processing is further interpreted by the auditory cortex to extract spectro-temporal modulation patterns (Shamma, 2003; Chih et al., 2005). An M-band modulation filterbank is employed in addition to the gammatone filterbank to model such functionality of the auditory cortex. By applying the modulation filterbank to each $\mathcal{H}_k(i,n)$, M outputs $\mathcal{H}_k(i,j,n)$ are generated where j denotes the jth modulation filter, $1 \leqslant j \leqslant M$. The filters in the modulation filterbank

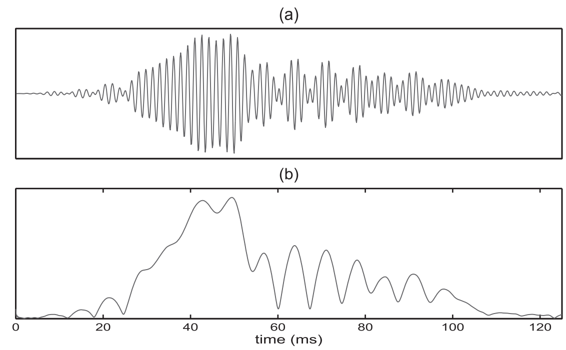

Fig. 3. Example of Hilbert envelope: (a) a 125 ms output of a critical-band filter centered at 650 Hz and (b) the corresponding Hilbert envelope.

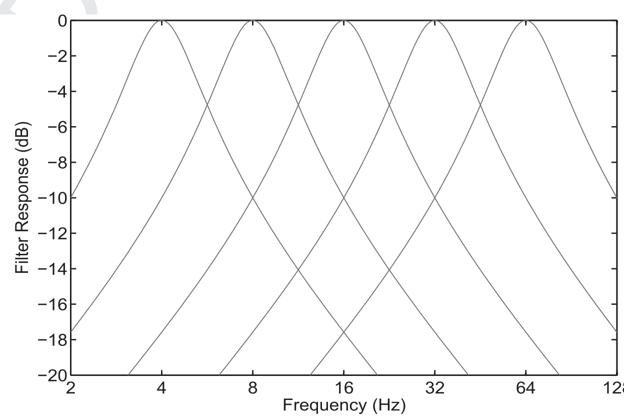

Fig. 4. Magnitude response of a 5-band modulation filterbank with center frequencies ranging from 4 to 64 Hz.

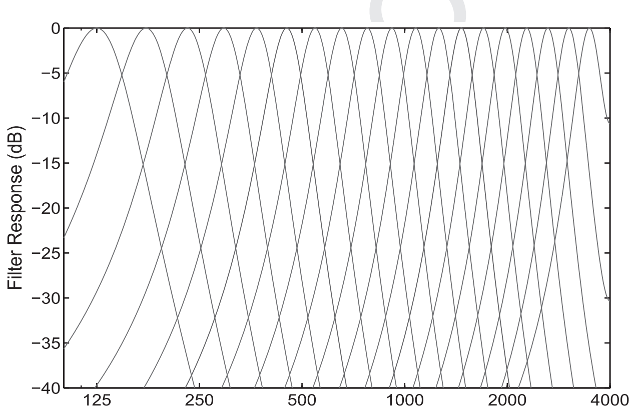

Fig. 2. Magnitude response of a 19-band auditory filterbank with center frequencies ranging from 125 Hz to 3.5 kHz.

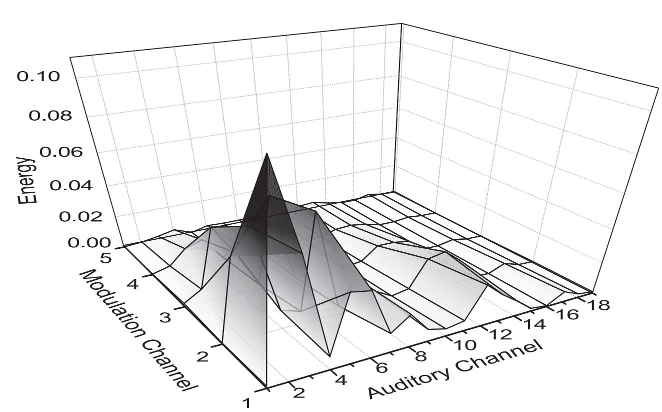

Fig. 5. $E_k(i,j)$ for one frame of a "*neutral*" speech file: low channel index indicates low frequency.

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

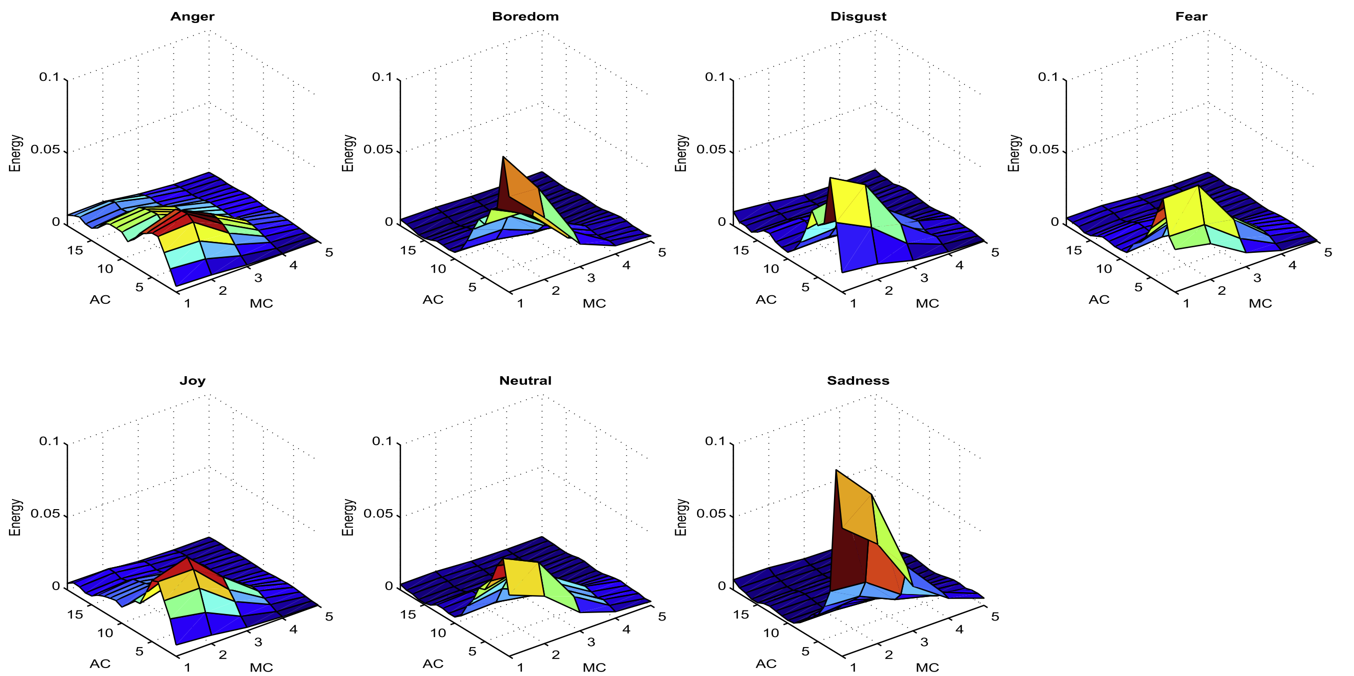

Fig. 6. Average $E(i,j)$ for seven emotion categories; for each emotion, $E_k(i,j)$ is averaged over all frames from all speakers of that emotion; "AC" and "MC" denote the acoustic and modulation frequency channels, respectively.

are second-order bandpass with quality factor set to 2, as suggested in (Ewert and Dau, 2000). In this work we use an $M = 5$ filterbank whose filter center frequencies are equally spaced on logarithm scale from 4 to 64 Hz. The filterbank was shown in preliminary experiments to strike a good balance between performance and model complexity. The magnitude response of the modulation filterbank is depicted in Fig. 4.

Lastly, the ST representation $E_k(i,j)$ of the kth frame is obtained by measuring the energy of $\mathcal{H}_k(i,j,n)$, given by:

$$E_k(i,j) = \sum_{n=1}^{L} |\mathcal{H}_k(i,j,n)|^2, \tag{4}$$

where $1 \leqslant k \leqslant T$ with L and T representing the number of samples in one frame and the total number of frames, respectively. For a fixed $j = j^*$, $E_k(i,j^*)$ relates the auditory spectral samples of modulation channel j^* after critical-band grouping. An example of $E_k(i,j)$ is illustrated in Fig. 5. By incorporating the auditory filterbank and the modulation filterbank, a richer two-dimensional frequency representation is produced and allows for analysis of modulation frequency content across different acoustic frequency channels.

Fig. 6 shows the ST representation $E(i,j)$ for the seven emotions in the Berlin database (cf. Section 4.1), where every $E(i,j)$ shown is the average over all the frames and speakers available in the database for an emotion. As illustrated in the figure, the average ST energy distribution over the joint acoustic-modulation frequency plane is similar for some emotions (e.g. *anger* vs. *joy*), suggesting they could become confusion pairs, while very distinct for some others (e.g. *anger* vs. *sadness*), suggesting they could be well discriminated from each other. As reasonably expected, the less expressive emotions such as *boredom* and *sadness* have significantly more low acoustic frequency energy than *anger* and *joy* (see also Fig. 7), corroborating findings in previous studies (Cowie et al., 2001; Scherer, 2003). The ST distribution for *neutral* peaks at 4 Hz modulation frequency, matching the nominal syllabic rate (Kanederaa et al., 1999). The peak shifts to a higher modulation frequency for *anger*, *joy*, and *fear*, suggesting a faster speaking rate for these emotions. Less expressive emotions such as *boredom* and *sadness* exhibit more prominently lowpass modulation spectral shapes, suggestive of lower speaking rates. Interestingly, *sadness* also shows increased energy for the last two modulation channels (centered at 32 and 64 Hz, respectively) relative to *anger* and *joy* (not evident from the plots, as the dominant energy is concentrated in

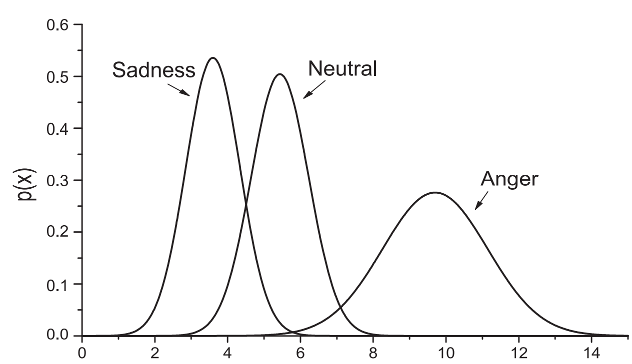

Fig. 7. Estimated pdfs of $\overline{\Phi}_3(1)$ for three basic emotions.

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

lower modulation channels and the absolute amount of energy at higher modulation frequencies is small). This might be due to the fact that sad speech is more breathy (Ishi et al., 2010), a phenomenon somewhat analogous to reverberant speech, whose effectively unvoiced excitation engenders more high modulation frequency energy (Falk and Chan, 2010b).

3. Feature extraction

In this section, we detail the proposed MSFs extracted from the ST representation. Short-term spectral features and prosodic features considered in our experiments are also described.

3.1. Modulation spectral features

Two types of MSFs are calculated from the ST representation, by means of spectral measures and linear prediction parameters. For each frame k, the ST representation $E_k(i,j)$ is scaled to unit energy before further computation, i.e. $\sum_{i,j} E_k(i,j) = 1$. Six spectral measures Φ_1–Φ_6 are then calculated on a per-frame basis. For frame k, $\Phi_{1,k}(j)$ is defined as the *mean* of the energy samples belonging to the jth modulation channel ($1 \leqslant j \leqslant 5$):

$$\Phi_{1,k}(j) = \frac{\sum_{i=1}^{N} E_k(i,j)}{N}. \tag{5}$$

Parameter Φ_1 characterizes the energy distribution of speech along the modulation frequency. The second spectral measure is the *spectral flatness* which is defined as the ratio of the geometric mean of a spectral energy measure to the arithmetic mean. In our calculation, $E_k(i,j)$ is used as the spectral energy measure at frame k for modulation band j and Φ_2 is thus defined as:

$$\Phi_{2,k}(j) = \frac{\sqrt[N]{\prod_{i=1}^{N} E_k(i,j)}}{\Phi_{1,k}(j)}. \tag{6}$$

A spectral flatness value close to 1 indicates a flat spectrum, while a value close to 0 suggests a spectrum with widely different spectral amplitudes. The third measure employed is the *spectral centroid* which provides a measure of the "center of mass" of the spectrum in each modulation channel. Parameter Φ_3 for the jth modulation channel is computed as:

$$\Phi_{3,k}(j) = \frac{\sum_{i=1}^{N} f(i) E_k(i,j)}{\sum_{i=1}^{N} E_k(i,j)}. \tag{7}$$

Two types of frequency measure $f(i)$ have been experimented: (1) $f(i)$ being the center frequency (in Hz) of the ith critical-band filter of the auditory filterbank and (2) $f(i)$ being the index of the ith criticalband filter, i.e., $f(i) = i$. No remarkable difference in performance is observed between the two measures, thus the latter is chosen for simplicity. Moreover, given the observation that adjacent modulation channels usually have considerable correlation, the spectral flatness and the centroid parameters of adjacent modulation channels also exhibit high correlation. In order to alleviate such information redundancy, $\Phi_{2,k}(j)$ and $\Phi_{3,k}(j)$ are only computed for $j \in \{1,3,5\}$.

Among the three aforementioned spectral measures, Φ_3 is observed to be particularly useful. Fig. 7 illustrates a representative example, where $\overline{\Phi}_3(1)$ is the average of $\Phi_{3,k}(1)$ computed over each utterance in the Berlin database belonging to three basic emotions: *anger*, *neutral*, and *sadness*, and the probability density function (PDF) of the averages for each emotion is estimated as a unimodal Gaussian. Considering *neutral* as a reference point, *anger* and *sadness* display an upward and downward shift of spectral centroid in acoustic frequency, respectively. This result is consistent with the ST patterns of these three emotions displayed in Fig. 6. Even though the PDFs of *sadness* and *neutral* overlap to some extent, good separation is shown for *anger* vs. *neutral*, and almost perfect discrimination is achieved between *anger* and *sadness*, using only one feature.

In addition to parameters that measure the spectral behavior of each individual modulation channel, additional spectral measures that measure the relationship of different modulation channels are computed. First, the 19 acoustic channels are grouped into four divisions: 1–4, 5–10, 11–15, and 16–19, namely D_l ($1 \leqslant l \leqslant 4$), which roughly correspond to frequency regions of <300 , 300–1000, 1000–2000, and >2000 Hz, respectively, and have been shown in pilot experiment to achieve a good compromise between the amount of fine details extracted from data and performance. Channels in the same division are summed: $\mathbb{E}_k(l,j) = \sum_{i \in D_l} E_k(i,j)$. Then the *modulation spectral centroid* (Φ_4) is calculated in a manner similar to Eq. 7:

$$\Phi_{4,k}(l) = \frac{\sum_{j=1}^{M} j \mathbb{E}_k(l,j)}{\sum_{j=1}^{M} \mathbb{E}_k(l,j)}. \tag{8}$$

Unlike $\Phi_{3,k}(j)$ which measures the spectral centroid in the acoustic frequency domain for modulation band j, $\Phi_{4,k}(l)$ calculates the centroid in the modulation frequency domain for D_l. The last two spectral measures $\Phi_{5,k}(l)$ and $\Phi_{6,k}(l)$ are the linear regression coefficient (slope) and the corresponding regression error (root mean squared error, RMSE) obtained by fitting a first-degree polynomial to $\mathbb{E}_k(l,j)$, $j = 1,\ldots,M$, in a least squares sense. By calculating Φ_4–Φ_6, information is extracted about the rate of change of the selected acoustic frequency regions, thereby compactly capturing the temporal dynamic cues. In total, 23 features are obtained from the ST representation per frame by applying the six spectral measures.

Besides taking the spectral measures described above, linear predication (LP) analysis is further applied to selected modulation channels j where $j \in \{1,3,5\}$, to extract the second set of MSFs from $E_k(i,j)$. This selection of modulation channels is also for the purpose of reducing information redundancy caused by high correlation between adjacent channels. The autocorrelation method for autoregressive (AR) modeling is used here. In order to suppress local details while preserving the broad structure beneficial to recognition, a 5th-order all-pole model is used to approximate the spectral samples. The computational cost of this AR modeling is negligible due to the low LP order and the small number of spectral samples per modulation channel (19 here). The LP coefficients obtained are further transformed into cepstral coefficients (LPCCs), and denoted as $C_k(n,j)$ $(0 \leqslant n \leqslant 5)$. The LPCCs have been shown to be generally more robust and reliable for speech recognition than the direct LP coefficients (Rabiner and Juang, 1993). We have tested both types and indeed the LPCCs yield better recognition performance in our application. Together with the 23 aforementioned features, a total of 41 MSFs are calculated frame-by-frame.

Although MSFs are extracted at a frame-level (FL), the common approach in current SER literature computes features at the utterance-level (UL) (e.g. Grimm et al., 2007a,b; Shami and Verhelst, 2007; Schuller et al., 2007b; Clavel et al., 2008; Lugger and Yang, 2008; Busso et al., 2009; Giannakopoulos et al., 2009; Sun and Torres, 2009), by applying descriptive functions (typically statistical) to the trajectories of FL features (and often also their derivatives to convey local dynamic information). The UL features capture the global properties and behaviors of their FL counterparts. It is desirable for the UL features to capture the supra-segmental characteristics of emotional speech that can be attributed to emotions rather than specific spoken-content and to effectively avoid problems such as spoken-content over-modeling (Vlasenko et al., 2007). Following the UL approach, two basic statistics: mean and standard deviation (std. dev.) of the FL MSFs are calculated in this work, producing 82 UL MSFs.

3.2. Short-term spectral features

3.2.1. MFCC features

The mel-frequency cepstral coefficients (MFCCs), first introduced in (Davis and Mermelstein, 1980) and successfully applied to automatic speech recognition, are popular short-term spectral features used for emotion recognition. They are extracted here for comparison with the proposed long-term MSFs. The speech signal is first filtered by a high-pass filter with a pre-emphasis coefficient of 0.97, and the first 13 MFCCs (including the zeroth order log-energy coefficient) are extracted from 25 ms Hamming-windowed speech frames every 10 ms. As a common practice, the delta and double-delta MFCCs describing local dynamics are calculated as well to form a 39-dimensional FL feature vector. The most frequently used UL MFCC features for emotion recognition include mean and std. dev. (or variance) of the first 13 MFCCs and their deltas (e.g. Grimm et al., 2007a,b; Schuller et al., 2007a; Vlasenko et al., 2007; Clavel et al., 2008; Lugger and Yang, 2008). In this work, we compute mean, std. dev., and 3rd–5th central moments of the first 13 MFCCs, as well as their deltas and double-deltas, giving 195 MFCC features in total. This MFCC feature set is an extension to the one used in (Grimm et al., 2007a, 2008), by further considering the delta coefficients.

3.2.2. PLP features

In addition to MFCCs, perceptual linear predictive (PLP) coefficients (Hermansky, 1990) are also extracted from speech, serving as an alternative choice of short-term spectral features for comparison. PLP analysis approximates the auditory spectrum of speech by an all-pole model that is more consistent with human hearing than conventional linear predictive analysis. A 5th-order model is employed as suggested in (Hermansky, 1990). The PLP coefficients are transformed to cepstral coefficients $c(n)$ $(0 \leqslant n \leqslant 5)$. The delta and double-delta coefficients are also considered. The aforementioned statistical parameters as used for MFCC are calculated for the PLP coefficients, giving 90 candidate PLP features.

3.3. Prosodic features

Prosodic features have been, among numerous acoustic features employed for SER, the most widely used feature type as mentioned in Section 1. Hence they are used here as a benchmark, and more importantly, to verify whether the MSFs can serve as useful additions to the extensively used prosodic features. The most commonly used prosodic features are based on pitch, intensity, and speaking rate. The features are estimated on a short-term frame basis, and their contours are used to compute UL features. The statistics of these trajectories are shown to be of fundamental importance for conveying emotional cues (Cowie et al., 2001; Nwe et al., 2003; Ververidis and Kotropoulos, 2006; Busso et al., 2009).

In total, 75 prosodic features are extracted as listed in Table 1. Note that a complete coverage of prosodic features is infeasible. Consequently, the features calculated here are by no means exhaustive, but serve as a representative sampling of the essential prosodic feature space. Pitch is computed for voiced speech using the pitch tracking algorithm in (Talkin, 1995), and intensity is measured for

Table 1
List of prosodic features.

Pitch, intensity, delta-pitch, delta-intensity
Mean, std. dev., skewness, kurtosis, shimmer, maximum, minimum, median, quartiles, range, differences between quartiles, linear & quadratic regression coefficients, regression error (RMSE)
Speaking rate
Mean and std. dev. of syllable durations
Ratio between the duration of voiced and unvoiced speech
Others
Zero-crossing rate (ZCR)
Teager energy operator (TEO)

active speech in dB. Note that shimmer is computed for the trajectories of pitch and intensity only, not their deltas. For a sequence x_n of length N_x, shimmer in this work is calculated as:

$$S = \frac{\frac{1}{N_x-1}\sum_{n=1}^{N_x-1}|x_n - x_{n+1}|}{\frac{1}{N_x}\sum_{n=1}^{N_x}x_n}. \tag{9}$$

Features related to speaking rate are also extracted using syllabic and voicing information as shown in the table. Moreover, the zero-crossing rate (ZCR) and the Teager energy operator (TEO) (Kaiser, 1990) of the speech signal are calculated as well, though they do not directly relate to prosody. TEO extracts useful information about the non-linear airflow structure of speech production (Zhou et al., 2001). The TEO for a discrete-time signal x_n is defined as:

$$\mathrm{TEO}(x_n) = x_n^2 - x_{n+1}x_{n-1}. \tag{10}$$

The mean Teager energy of the speech signal is used as a feature, and is found to be an effective feature in our experiments (cf. Section 5.1.3).

4. Emotional speech data

4.1. Berlin emotional speech database

The Berlin emotional speech database (Burkhardt et al., 2005) is used for experiments classifying discrete emotions. This publicly available database is one of the most popular databases used for emotion recognition, thus facilitating comparisons with other works. Ten actors 5m/5f) each uttered 10 everyday sentences (five short and five long, typically between 1.5 and 4 s) in German, sentences that can be interpreted in all of seven emotions acted. The raw database (prior to screening) has approximately 800 sentences and is further evaluated by a subjective perception test with 20 listeners. Utterances scoring higher than 80% emotion recognition rate and considered natural by more than 60% listeners are included in the final database. The numbers of speech files for the seven emotion categories in the screened Berlin database are: *anger* (127), *boredom* (81), *disgust* (46), *fear* (69), *joy* (71), *neutral* (79) and *sadness* (62).

4.2. Vera am Mittag (VAM) database

The VAM database (Grimm et al., 2008) is a relatively new database containing spontaneous emotions, created within a three-dimensional emotion space framework. It was recorded from a German TV talk-show "Vera am Mittag". There are three individual modules in the complete VAM database: VAM-Audio, VAM-Video, and VAM-Faces, containing audio signal only, audio + visual signals, and face images, respectively. In this work, only the VAM-Audio module is employed and hereinafter it is simply referred to as the VAM database. The recordings are manually segmented at the utterance level. Emotions in the database are described in an emotion space consisting of three emotion primitives: *valence* (or *evaluation*, ranging from negative to positive), *activation* (with levels from low to high) and *dominance* (the apparent strength of the speaker, i.e. ability to handle a situation) (Grimm et al., 2007a). The continuous emotion values on each primitive scale are obtained through subjective assessment. The VAM database contains two parts: VAM I with 478 utterances from 19 speakers (4m/15f) and 17 human evaluators assessing the primitives, and VAM II with 469 utterances from 28 speakers (7m/21f) and six evaluators. The distributions of the three primitives in the VAM database (I + II) are shown in Fig. 8 (Grimm et al., 2008). It can be seen from the histograms that compared to *activation* and *dominance*, the distribution of *valence* is less balanced. The database contains mostly neutral and negative emotions, due to the topics discussed in the talk-show (Grimm et al., 2008).

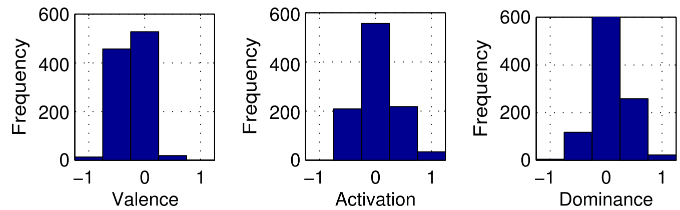

Fig. 8. Distribution of the emotion primitives in the overall VAM database (Grimm et al., 2008).

5. Experiments

In this section, results of the experimental evaluation are presented. Support vector machines (SVMs) (Vapnik, 1995) are used for recognition of both discrete and continuous emotions. While support vector classification finds the separation hyperplane that maximizes the margin between two classes, support vector regression determines the regression hyperplane that approximates most data points with ϵ precision. The SVM implementation in (Chang and Lin, 2009) is adopted with the radial basis function (RBF) kernel employed. The design parameters of SVM are selected using training data via a grid search on a base-2 logarithmic scale. In general, the RBF kernel can be a good choice as justified in (Hsu et al., 2007) because: (1) it can model the non-linear relation between attributes and target values well; (2) the linear kernel is a special case of RBF kernel; (3) it has less hyperparameters than the polynomial kernel; and (4) it has less numerical difficulties compared to polynomial and sigmoid kernels. Features from training data are linearly scaled to $[-1, 1]$ before applying SVM, with features from test data scaled using the trained linear mapping function as suggested in (Hsu et al., 2007).

5.1. Experiment I: discrete emotion classification

All results achieved on the Berlin database are produced using 10-fold cross-validation. The data partitioning is

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

based on random sampling of files from a pool wherein all speakers are mixed; hence the cross-validation is not speaker-independent. Moreover, samples in each class have been randomly divided into 10 disjoint subsets approximately equal in size. Each validation trial takes nine subsets from every class for training, with the remaining subset kept unseen from the training phase and used for testing only.

A two-stage feature selection scheme is employed and is described in Section 5.1.1. The proposed MSFs are initially compared to MFCC and PLP features in Section 5.1.2, with their contribution as supplementary features to prosodic features investigated in Section 5.1.3. The effect of taking a speaker normalization (SN) step to pre-process the data prior to data partitioning for cross-validation is also investigated, where features are first mean and variance normalized within the scope of each speaker to compensate for speaker variations, as performed in (Vlasenko et al., 2007). Let $f_{u,v}(n)$ $(1 \leqslant n \leqslant N_{u,v})$ stand for the uth feature from speaker v with $N_{u,v}$ denoting its sample size which in our case, is the number of all available samples in the database from that speaker. Then the new feature $f'_{u,v}(n)$ processed by SN is given by:

$$f'_{u,v}(n) = \frac{f_{u,v}(n) - \overline{f_{u,v}}}{\sqrt{\frac{1}{N_{u,v}-1}\sum_{m=1}^{N_{u,v}}\left(f_{u,v}(m) - \overline{f_{u,v}}\right)^2}}, \tag{11}$$

where $\overline{f_{u,v}} = \frac{1}{N_{u,v}}\sum_{n=1}^{N_{u,v}} f_{u,v}(n)$. Since SN requires the system to know the speaker in advance, it might not be realistic in many applications. Nevertheless its results could still be interesting to see, hence included in following experiments.

5.1.1. Feature selection

Using all the features for machine learning might deteriorate recognition performance due to the curse of dimensionality (Bishop, 2006). To this end, a two-stage feature selection scheme is proposed to reduce the number of features. The first stage calculates the Fisher discriminant ratio (FDR) to rank each feature individually, which can quickly eliminate irrelevant ("noisy") features. The normalized multi-class FDR for the uth feature is defined as:

$$\mathrm{FDR}(u) = \frac{2}{C(C-1)}\sum_{c_1}\sum_{c_2}\frac{\left(\mu_{c_1,u} - \mu_{c_2,u}\right)^2}{\sigma^2_{c_1,u} + \sigma^2_{c_2,u}}, \tag{12}$$

with $1 \leqslant c_1 < c_2 \leqslant C$, where $\mu_{c_1,u}$ and $\sigma^2_{c_1,u}$ are mean and variance of the uth feature for the c_1th class, and C is the total number of classes. This FDR measure is normalized by the number of binary comparisons made between two classes. The measure favors features with well-separated means across classes and small within-class variances. Features of little discrimination power can then be removed by FDR thresholding. Here the threshold is empirically set to 0.15 as further increasing the threshold results in no improvement of performance, which reduces roughly 10% of the proposed features and 15% of prosodic features, but up to 50% of MFCC features and 40% of PLP features. Thus the FDR step is important for the short-term spectral feature pools, which would otherwise be rather noisy for feature mining.

In the second stage, two techniques are experimented to obtain good features from the pre-screened feature pools for SVM classification. The first technique is sequential forward feature selection (SFS) (Kittler, 1978), which iteratively augments the selected feature subset and considers the combined effect of features and SVM classifier during the evaluation process. The SFS algorithm also helps to visualize changes of recognition accuracy as the selected feature subset evolves, and thus provides a straightforward way to compare performance of different feature combinations.

The second technique is the well-known multi-class linear discriminant analysis (LDA) (Bishop, 2006). It finds the transformation optimizing the Fisher objective, that in turn maximizes the between-class distance and minimizes the within-class distance simultaneously. LDA can offer a drastic reduction in feature dimensionality for high-dimensional data and effectively alleviates the curse of dimensionality, hence particularly useful in practice if the size of the training data is limited relative to the number of features. The main limitation of LDA is that it cannot be applied to regression problems. LDA solves the generalized eigen-problem:

$$\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}, \tag{13}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between-class and within-class scatter matrices, respectively. The eigenvectors $\mathbf{w}$ are used to form the columns of the transformation matrix $\mathbf{W}$ which transforms data point $\mathbf{x}$ to $\mathbf{y} = \mathbf{W}^T\mathbf{x}$. The components of $\mathbf{y}$ constitute the LDA-transformed features. Since the maximum rank of $\mathbf{S}_b$ for a C-class problem is $C-1$, the maximum number of LDA features is also $C-1$. For our seven-class case, all six LDA-transformed features are used to design SVM classifiers.

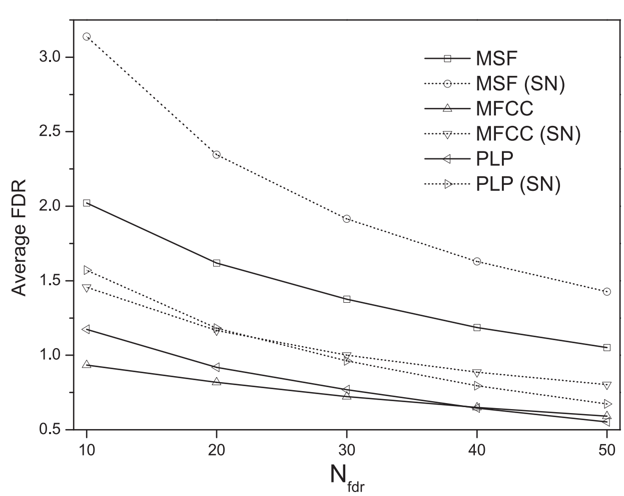

Fig. 9. Average FDR curves for the proposed, MFCC and PLP features (Berlin).

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

Table 2
Recognition results for MSF, MFCC, and PLP features (Berlin); boldface indicates the best performance in each test.

Test	Feature	Method	SN	Recognition rate (%)							Average
				Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	
#1	MSF (31/74)	SFS	No	**92.1**	**86.4**	**73.9**	62.3	49.3	**83.5**	**98.4**	**79.6**
	MFCC (49/92)			91.3	80.3	67.4	**65.2**	**50.7**	79.8	85.5	76.5
	PLP (48/51)			89.8	79.0	50.0	58.0	43.7	72.2	83.9	71.2
#2	MSF	LDA	No	**91.3**	**86.4**	**78.3**	71.0	**60.6**	**83.5**	**88.7**	**81.3**
	MFCC			83.5	85.2	**78.3**	**76.8**	53.5	79.8	83.9	77.9
	PLP			88.2	74.1	56.5	55.1	49.3	77.2	80.7	71.4
#3	MSF (41/73)	SFS	Yes	89.8	**88.9**	67.4	**81.2**	**59.2**	**84.8**	**98.4**	**82.8**
	MFCC (37/96)			88.2	82.7	**71.7**	76.8	54.9	76.0	90.3	78.5
	PLP (26/51)			**90.6**	72.8	56.5	66.7	46.5	81.0	95.2	75.1
#4	MSF	LDA	Yes	**90.6**	**87.7**	76.1	**91.3**	**70.4**	**84.8**	**91.9**	**85.6**
	MFCC			85.0	79.0	**78.3**	81.2	54.9	79.8	88.7	78.7
	PLP			**90.6**	80.3	58.7	72.5	56.3	81.0	88.7	77.8

5.1.2. Comparison with short-term spectral features

The proposed features are first compared to two short-term spectral feature types: MFCC and PLP features, by means of FDR scores before applied to SVMs. The features are ranked by their FDR values (calculated using all instances in the Berlin database before data partitioning). Curves depicting the FDR values averaged over the top N_{fdr} FDR-ranked features are shown in Fig. 9 as a function of N_{fdr}. These average FDR curves can be viewed as rough indicators of discrimination power of the three feature types independent of a specific classifier. As depicted in the figure, the MSFs consistently exhibit considerably better discrimination power than the other two spectral feature types, whose FDR curves are relatively close. Speaker normalization is shown to be beneficial, as it boosts all the FDR curves. However, selecting features based solely on FDR can be hazardous, as it only evaluates features individually. Effects such as feature correlation and classifier properties have not been taken into account. Therefore, a subsequent step to find better feature combinations is necessary, as performed by the second stage of our feature selection scheme.

The numeric results of applying SFS and LDA techniques to the FDR screened feature pools are detailed in Table 2 with SVMs employed for classification and the results averaged over the 10 cross-validation trials. The average recognition rate is measured as the number of samples from all emotions correctly recognized divided by the total number of samples. For SFS, the results shown in Table 2 are for the number of features that yields the best average performance. This best-performing number of features selected by SFS and the feature pool size (after FDR screening) are given by "X" and "Y", respectively, in the format of "X/Y" as shown in the table. Because the pre-screened PLP feature pool consists of 51 features only, SFS is terminated at 50 features for all feature types so that, for fair comparison, the maximum number of features that can be selected from each feature type is the same. For LDA, all the six transformed features are used as mentioned in Section 5.1.1.

As shown in Table 2, the proposed MSFs achieve the best accuracy in most emotions and in overall performance, reaching up to 85.6% average recognition rate (using LDA with SN). Applying SN improves both SFS and LDA performance for all features. It is also interesting to see that in these tests, using six LDA transformed features delivers even higher accuracy than using dozens of SFS features, indicating that the effective reduction of feature dimensionality offered by LDA indeed contributes to recognition performance. The average recognition rate for the three feature types are further depicted in Fig. 10, as a function of the number of features selected by SFS. It is clear from Fig. 10 that the MSFs consistently outperform MFCC and PLP features, irrespective of SN, though performing SN boosts the accuracy curves more notably for short-term spectral features than for the MSFs. Moreover, as indicated by the figure, MFCC features appear to be more suitable for emotion recognition relative to PLP features, despite FDR results suggesting the two feature types to be comparable. Such finding resonates with the aforementioned fact that additional procedures are needed to explore feature combinations after FDR pre-screening.

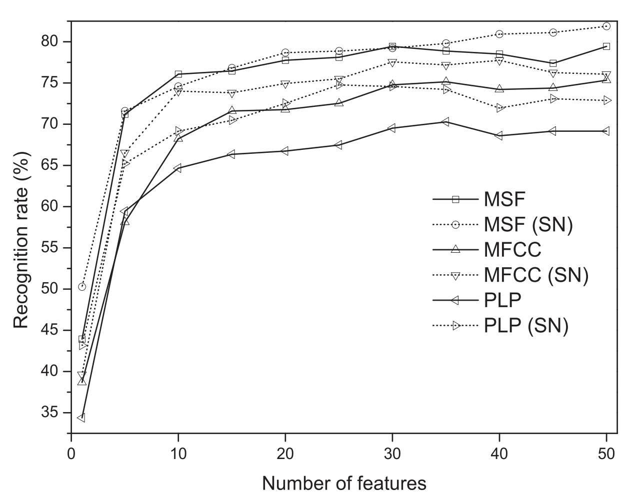

Fig. 10. Recognition results for MSF, MFCC, and PLP features selected by SFS (Berlin).

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

5.1.3. Comparison with prosodic features

Due to the widespread use of prosodic features, it is important to study the contribution of the proposed MSFs as complementary features. Comparisons are hence performed using: (1) prosodic features only and (2) combined prosodic and proposed features. Although the prosodic features extracted in our study do not exhaust all possibilities, they do cover important aspects of prosodic information. Consequently, recognition results for this typical prosodic feature set can serve as a rule of thumb for the underlying extensive prosodic feature space.

The average FDR curves of the two feature types and their combinations are shown in Fig. 11, where the label "PROS" stands for prosodic features. As can be seen from the figure, the proposed features outperform prosodic features in terms of FDR scores, and the discrimination power of the prosodic feature pool can be substantially enhanced after the inclusion of MSFs. Both feature types benefit considerably from SN. The recognition rate trajectories using features selected by SFS are illustrated in Fig. 12, where the contribution of the MSFs is evident.

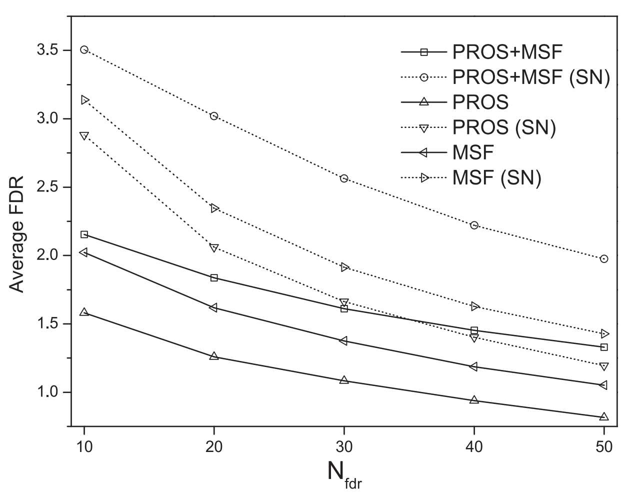

Fig. 11. Average FDR curves for prosodic features, MSFs, and their combination (Berlin).

Table 3
Top 10 features for prosodic, proposed, and combined features as ranked by AFR (Berlin).

Rank	Feature	AFR	Rank	Feature	AFR
Prosodic features					
1	TEO	2.4	6	$Q_3 - Q_2$ of pitch	8.4
2	Mean syllable duration	3.3	7	Kurtosis of intensity	8.6
2	$Q_3 - Q_2$ of delta-pitch	3.3	8	Q_3 of delta-pitch	9.1
4	Slope of pitch	7.3	9	Minimum of intensity	9.8
5	Q_1 of delta-pitch	8.3	9	ZCR	9.8
Proposed features					
1	Mean of $\Phi_{1,k}(3)$	2.3	6	Mean of $\Phi_{6,k}(2)$	7.0
2	Mean of $C_k(4,5)$	5.0	7	Mean of $\Phi_{1,k}(2)$	7.1
3	Mean of $\Phi_{3,k}(1)$	5.6	8	Mean of $\Phi_{4,k}(4)$	7.7
4	Mean of $\Phi_{3,k}(3)$	6.0	9	Mean of $\Phi_{3,k}(5)$	8.7
5	Mean of $\Phi_{5,k}(2)$	6.4	10	Mean of $C_k(1,5)$	9.1
Combined features					
1	Mean of $\Phi_{1,k}(3)$	3.0	6	Mean of $\Phi_{6,k}(2)$	7.8
2	Mean syllable duration	4.5	7	Q_1 of pitch	9.0
3	Mean of $\Phi_{3,k}(3)$	6.0	8	Q_1 of delta-pitch	9.1
4	Mean of $\Phi_{3,k}(1)$	6.1	9	Slope of pitch	9.2
5	Mean of $\Phi_{5,k}(2)$	7.2	10	Mean of $C_k(4,5)$	9.4

The top features (no SN) selected by SFS from prosodic features, MSFs, and their combination are presented in Table 3, by means of average feature rank (AFR). Denote the candidate feature pool as F. The AFR of the uth feature $f_u \in F$ given R cross-validation trials is calculated as:

$$\mathrm{AFR}(f_u) = \frac{1}{R}\sum_{r=1}^{R} \text{rank of } f_u \text{ in the } r\text{th trial.} \qquad (14)$$

If f_u is not selected in a trial, its rank is replaced by a penalty value P. The AFR tables here are produced by

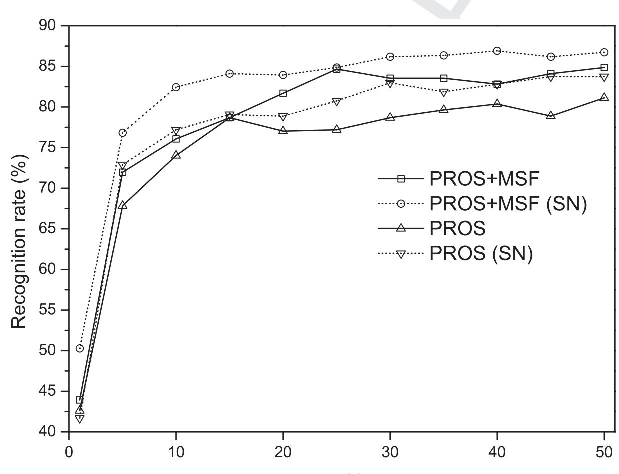

Fig. 12. Comparison between prosodic features and their combination with MSFs (Berlin).

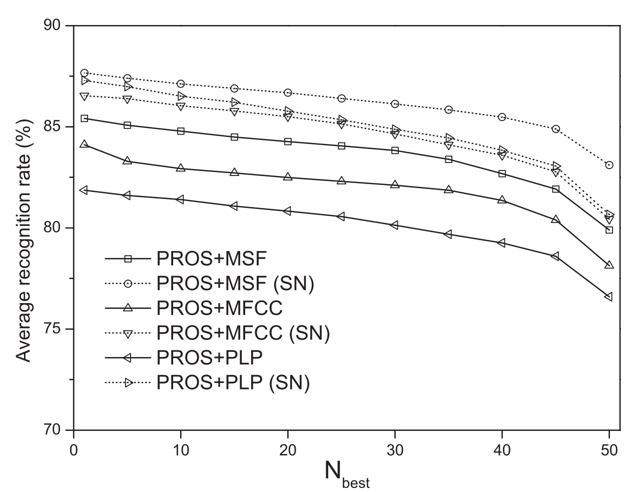

Fig. 13. Comparison between different combinations of spectral and prosodic features (Berlin).

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

selecting the top 10 features in each trial with the penalty value P set to 11. Features with small AFR values, such as TEO and $\Phi_{1,k}(3)$, are consistently top ranked across the cross-validation trials. In practice, such results can be used to pick features when forming a final feature set to train the classifier. We also compiled AFR tables for features with SN (not shown). It is observed that nearly half of the features in Table 3 appear in the SN case as well, though with different ranks, and several features remain top ranked regardless of SN, such as TEO, mean syllable duration, $\Phi_{1,k}(3)$ and $\Phi_{3,k}(1)$.

The short-term spectral features are also considered for comparison. SFS results with different combinations of spectral feature type and prosodic features are given in Fig. 13, where the proposed features achieve the highest recognition rate, either with SN (MSF: 87.7%, MFCC: 86.5%, PLP: 87.3%) or without SN (MSF: 85.4%, MFCC: 84.1%, PLP: 81.9%). Although the best performance of different feature combinations might seem close, the advantage of the proposed features is most notable in Fig. 14. The label N_{best} denotes that the N_{best} ($1 \leqslant N_{best} \leqslant 50$) highest recognition rates (among the 50 recognition rates obtained by selecting 1–50 features using SFS) are chosen, with their average being calculated and depicted in the figure. As shown in Fig. 14, the MSFs always furnish the highest average recognition rate relative to MFCC and PLP features, irrespective of SN. But among the three spectral feature types evaluated by SFS, PLP features (combined with prosodic features) turn out to receive the largest performance gain from SN.

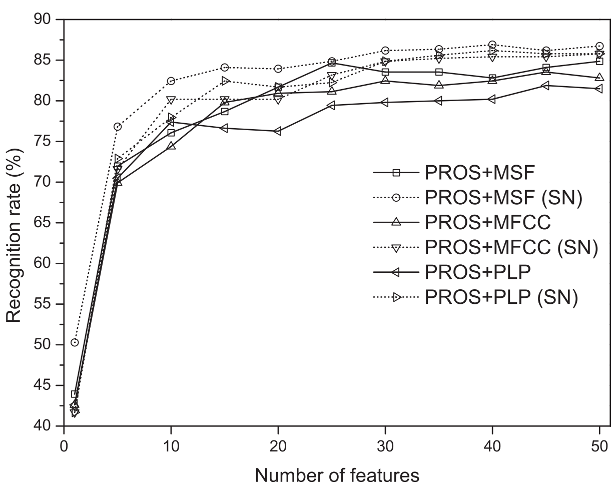

Fig. 14. Average recognition performance of N_{best} for different combinations of spectral and prosodic features (Berlin).

Table 4 details the results achieved by the combined features, with recognition rate shown for each emotion. LDA results are included as well. The row labeled "↓ %" indicates the percentage reduction of error rate obtained by adding the MSFs to prosodic features, which is calculated as:

$$\downarrow\ \% = \frac{\mathrm{RR}_{\mathrm{PROS+MSF}} - \mathrm{RR}_{\mathrm{PROS}}}{1 - \mathrm{RR}_{\mathrm{PROS}}} \times 100\%, \tag{15}$$

where "RR" represents recognition rate. Similar to the case where spectral features are evaluated individually, the MSFs achieve the highest overall accuracy when combined with prosodic features, and up to 91.6% recognition rate can be obtained using LDA with SN. Applying LDA with SN also gives the best recognition performance for

Table 4
Recognition results for prosodic and combined features (Berlin); boldface indicates the best performance in each test.

Test	Feature	Method	SN	Recognition rate (%)							Average
				Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	
#1	PROS (50/65)	SFS	No	89.8	87.7	73.9	84.1	**59.2**	79.8	83.9	81.1
	PROS + MSF (48/139)			**94.5**	**88.9**	**76.1**	84.1	57.8	**89.9**	**96.8**	**85.4**
	PROS + MFCC (38/157)			93.7	**88.9**	71.7	**85.5**	56.3	**89.9**	90.3	84.1
	PROS + PLP (45/116)			90.6	84.0	71.7	78.3	**59.2**	**89.9**	88.7	81.9
	↓ %			46.1	9.8	8.4	0.0	−3.4	50.0	80.1	22.8
#2	PROS	LDA	No	87.4	82.7	78.3	79.7	49.3	82.3	83.9	78.7
	PROS + MSF			**90.6**	**92.6**	87.0	**82.6**	**62.0**	87.3	88.7	**85.0**
	PROS + MFCC			84.3	87.7	**93.5**	79.7	**62.0**	**88.6**	**91.9**	83.6
	PROS + PLP			87.4	90.1	89.1	68.1	57.8	84.8	88.7	81.3
	↓ %			25.4	57.2	40.1	14.3	25.1	28.3	29.8	29.6
#3	PROS (41/64)	SFS	Yes	92.9	91.4	78.3	81.2	56.3	87.3	90.3	83.9
	PROS + MSF (48/137)			94.5	87.7	**82.6**	**84.1**	63.4	**96.2**	**98.4**	**87.7**
	PROS + MFCC (44/160)			93.7	**92.6**	**82.6**	**84.1**	**66.2**	84.8	95.2	86.5
	PROS + PLP (44/115)			**96.1**	**92.6**	78.3	82.6	63.4	92.4	95.2	87.3
	↓ %			22.5	−43.0	19.8	15.4	16.3	70.1	83.5	23.6
#4	PROS	LDA	Yes	89.8	85.2	80.4	87.0	62.0	93.7	93.6	85.2
	PROS + MSF			**93.7**	**96.3**	**91.3**	**89.9**	**73.2**	**94.9**	**100**	**91.6**
	PROS + MFCC			88.2	87.7	89.1	84.1	69.0	89.9	91.9	85.8
	PROS + PLP			85.8	87.7	78.3	84.1	67.6	93.7	91.9	84.7
	↓ %			38.2	75.0	55.6	22.3	29.5	19.0	100	43.2

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

Table 5
Confusion matrix for using only prosodic features (Berlin).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate (%)
Anger	**114**	0	1	3	8	1	0	89.8
Boredom	0	**69**	1	0	0	8	3	85.2
Disgust	0	1	**37**	2	0	5	1	80.4
Fear	4	0	0	**60**	3	2	0	87.0
Joy	19	0	1	2	**44**	5	0	62.0
Neutral	1	2	1	0	0	**74**	1	93.7
Sadness	0	3	0	0	0	1	**58**	93.6
Precision (%)	82.6	92.0	90.2	89.6	80.0	77.1	92.1	

Table 6
Confusion matrix for using prosodic and proposed features (Berlin).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Rate (%)
Anger	**119**	0	1	1	6	0	0	93.7
Boredom	0	**78**	0	0	0	3	0	96.3
Disgust	0	0	**42**	1	1	2	0	91.3
Fear	2	0	0	**62**	3	2	0	89.9
Joy	12	0	1	3	**52**	3	0	73.2
Neutral	0	3	1	0	0	**75**	0	94.9
Sadness	0	0	0	0	0	0	**62**	100
Precision (%)	89.5	96.3	93.3	92.5	83.9	88.2	100	

prosodic features (85.2%). For MFCC and PLP features, however, superior results are observed for SFS with SN.

Two confusion matrices are shown in Tables 5 and 6 (left-most column being the true emotions), for the best recognition performance achieved by prosodic features alone and combined prosodic and proposed features (LDA + SN), respectively. The *rate* column lists per class recognition rates, and *precision* for a class is the number of samples correctly classified divided by the total number of samples classified to the class. We can see from the confusion matrices that adding MSFs contributes to improving the recognition and precision rates of all emotion categories. It is also shown that most emotions can be correctly recognized with above 89% accuracy, with the exception of *joy*, which forms the most notable confusion pair with *anger*, though they are of opposite *valence* in the activation–valence emotion space (Cowie et al., 2001). This might be due to the fact that *activation* is more easily recognized by machine than *valence*, as indicated by the regression results for the emotion primitives on the VAM database presented in Section 5.2.

As aforementioned, the cross-validation scheme used so far is not entirely speaker-independent. We further investigate the effect of speaker dependency for SER by doing "leave-one-speaker-out" (LOSO) cross-validation, wherein the training set does not contain a single instance of the speaker in the test set. LOSO results with different features are presented in Table 7 and compared with the previous speaker-dependent 10-fold results in Tables 2 and 4 (simply denoted as "10-fold" in the table). As shown in the table, emotion recognition accuracy under the more stringent LOSO condition is lower than when test speakers are represented in the training set. This expected behavior applies to all feature types. When tested alone, the MSFs clearly outperform MFCC and PLP features. When combined with prosodic features, MFCC features yield the best performance if SN is applied; otherwise, the MSFs still prevail. As SN is not applied in typical real-life applications, the proposed features might be more suitable for these more realistic scenarios.

However, it should also be noted that the Berlin database has limited phonetic content (10 acted sentences), hence limiting the generalizability of the obtained results.

Table 7
Recognition results for LOSO cross-validation (Berlin).

Feature	Recognition rate (%)			
	SFS		LDA	
	10-fold	LOSO	10-fold	LOSO
MSF	79.6	74.0	81.3	71.8
MFCC	76.5	65.6	77.9	66.0
PLP	71.2	63.2	71.4	65.0
MSF (SN)	82.8	78.1	85.6	79.1
MFCC (SN)	78.5	71.8	78.7	72.3
PLP (SN)	75.1	72.3	77.8	72.5
PROS	81.1	75.0	78.7	71.2
PROS + MSF	85.4	78.1	85.0	76.3
PROS + MFCC	84.1	75.1	83.6	75.5
PROS + PLP	81.9	75.3	81.3	72.1
PROS (SN)	83.9	83.0	85.2	80.2
PROS + MSF (SN)	87.7	83.2	91.6	80.9
PROS + MFCC (SN)	86.5	85.8	85.8	82.4
PROS + PLP (SN)	87.3	84.3	84.7	78.7

It is also useful to briefly review performance figures reported on the Berlin database by other works. Although the numbers cannot be directly compared due to factors such as different data partitioning, they are still useful for general benchmarking. Unless otherwise specified, the results cited here are achieved for recognizing all seven emotions with UL features. For works that do not use speaker normalization, 86.7% recognition rate is achieved under 10-fold cross-validation in (Schuller et al., 2006), by using around 4000 features. The accuracy is slightly improved to 86.9% after optimizing the feature space, but the dimensionality of the optimized space is not reported. In (Lugger and Yang, 2008), 88.8% accuracy is achieved by employing a three-stage classification scheme, but based on recognition of six emotions only (no *disgust*). Among the works that use speaker normalization, 83.2% recognition rate is obtained in a leave-one-speaker-out experiment (Vlasenko et al., 2007) by extracting around 1400 acoustic features for data mining. However, no information is provided about the final number of features used. The accuracy is further improved to 89.9% by integrating both UL and FL features.

5.2. Experiment II: continuous emotion regression

The well-established descriptive framework that uses discrete emotions offers intuitive emotion descriptions and is commonly used. The combination of basic emotion categories can also serve as a convenient representation of the universal emotion space (Ekman, 1999). However, recent research efforts also show an increasing interest in dimensional representations of emotions for SER (Grimm et al., 2007a; Grimm et al., 2007b; Wollmer et al., 2008; Giannakopoulos et al., 2009). The term "dimensional" here refers to a set of primary emotion attributes that can be treated as the bases of a multi-dimensional emotion space, wherein categorical descriptors can be situated by coordinates. A dimensional framework allows for gradual change within the same emotion as well as transition between emotional states. In this experiment, we recognize the three continuous emotion primitives – *valence*, *activation* and *dominance* – in the VAM database.

5.2.1. Continuous emotion recognition

Leave-one-out (LOO) cross-validation is used to enable comparisons with the results reported in (Grimm et al., 2007a). In this LOO test, the speaker in the test instance is also represented in the training set. Akin to the comparison framework in Section 5.1, regression is performed using (1) spectral features, (2) prosodic features, and (3) combined prosodic and spectral features (without SN). The SFS algorithm is employed to select the best features for the support vector regressors (SVRs). Experiments are carried out on three datasets: VAM I, VAM II, and VAM I + II. Ideally, LOO cross-validation on N-sample data requires SFS to be applied to all the N different training sets, each containing $N-1$ samples. However, since N is reasonably large here (478, 469, and 947 for VAM I, II, and I + II, respectively), including the test sample hardly impacts the SFS selections. Thus in each experiment, feature selection is carried out using all the samples of the corresponding dataset. LOO is then used to re-train the regressor using the $N-1$ samples in each validation trial and test the remaining sample.

Correlation coefficient r and mean absolute error e are the two performance measures to be calculated. For two sequences x_n and y_n of the same length, the correlation coefficient is calculated using Pearson's formula:

$$r = \frac{\sum_n (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_n (x_n - \bar{x})^2 \sum_n (y_n - \bar{y})^2}}, \qquad (16)$$

where $\bar{x}(\bar{y})$ is the average of x_n (y_n). Mean absolute error is used, identical to the error measure employed in (Grimm et al., 2007a).

Mean correlations, averaged over the three primitives, are shown in Fig. 15 for the spectral features. Unlike the emotion classification results for the Berlin database, MFCC features furnish better regression performance on the VAM datasets as suggested by the figure. A closer examination (see also Table 8) reveals that the gain of MFCC features over MSFs is mainly due to the former's better performance on the *valence* primitive, especially for the VAM II dataset. Such results suggest that MFCC features are more competent than the proposed features at

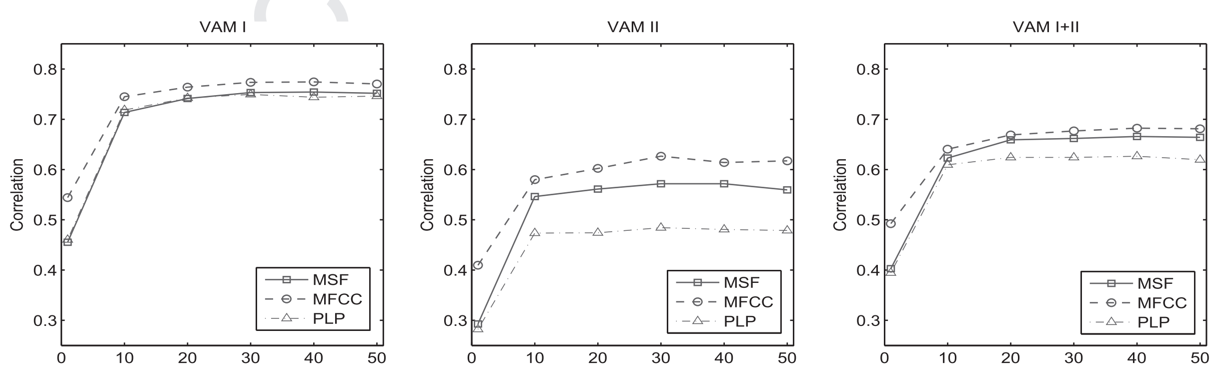

Fig. 15. Mean correlations for MSF, MFCC and PLP features selected using SFS (VAM).

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

Table 8
Regression results for continuous emotions on the VAM database using MSF, MFCC, and PLP features.

Dataset	Feature	Correlation (r)			Absolute error (e)			Average	
		Valence	Activation	Dominance	Valence	Activation	Dominance	$\bar{r}$	$\bar{e}$
VAM I	MSF	0.60	0.86	0.81	0.11	0.15	0.15	0.76	0.14
	MFCC	0.65	0.87	0.81	0.11	0.14	0.15	0.78	0.13
	PLP	0.61	0.86	0.79	0.11	0.16	0.16	0.75	0.14
VAM II	MSF	0.32	0.74	0.66	0.15	0.16	0.16	0.57	0.16
	MFCC	0.46	0.74	0.68	0.14	0.16	0.15	0.63	0.15
	PLP	0.26	0.63	0.59	0.15	0.18	0.16	0.49	0.16
VAM I + II	MSF	0.46	0.80	0.75	0.13	0.17	0.16	0.67	0.15
	MFCC	0.52	0.79	0.74	0.13	0.17	0.16	0.68	0.15
	PLP	0.42	0.75	0.70	0.13	0.18	0.17	0.62	0.16

determining the positive or negative significance of emotions (i.e. *valence*). On the other hand, the two feature types provide very close performance for recognizing *activation* and *dominance* as indicated by the correlation curves for individual primitives (not shown). PLP features give inferior regression results on VAM II and VAM I + II, again becoming the spectral feature type that yields the worst SER outcomes. The highest mean correlation achieved by each feature type in Fig. 15 is further interpreted in Table 8 by showing the regression performance for individual primitives. Comparing the three primitives, *activation* receives the best correlations (0.63–0.87), while *valence* shows significantly lower correlations (0.26–0.65) for all features. It is also observed that the absolute regression errors between the different spectral feature types are quite close.

Regression of prosodic and combined features is considered in Fig. 16 and Table 9. The machine recognition and human subjective evaluation results given in (Grimm et al., 2007a) are also included in Table 9 for reference. Note that in (Grimm et al., 2007a), only the standard deviation of subjective scores is presented for each primitive. The error is the standard deviation minus a constant bias term that depends on the number of evaluators, which can be inferred from the paper. As shown in the figure and table, adding spectral features to prosodic features improves the correlations for all primitives on all datasets, and slightly reduces the estimation errors in some cases. The MFCC features (combined with prosodic features) still deliver the best performance, followed by the proposed features. Again the advantage of the MFCC features over MSFs is on recognizing *valence*. The performance gaps between MSFs and MFCC features in Table 8 appear narrowed in Table 9, indicating that the prosodic features enhance MSF performance more than MFCC performance, mainly because the prosodic features contribute more to improving *valence* recognition for the MSF case, even though overall MFCC performance is slightly better.

The recognition tendency for the primitives in Table 9 is the same as the trend observed in Table 8. The primitive *activation* is best estimated with up to 0.90 correlation achieved on VAM I, followed by *dominance* whose regression performance is more moderate. Even though the inclusion of spectral features improves the correlations for *valence*, the attained values are relatively low for all feature combinations. Nevertheless, even human evaluations give poor correlations for recognizing *valence* compared to the other two primitives.

Overall, the recognition system using combined features yields higher correlations and smaller estimation errors compared to the machine recognition results in (Grimm et al., 2007a). The performance of the proposed recognition system appears to be somewhat superior to human assessment; this however merely reflects the capability of machines to be more consistent (but not more accurate) than humans in performing the emotion labeling task. In

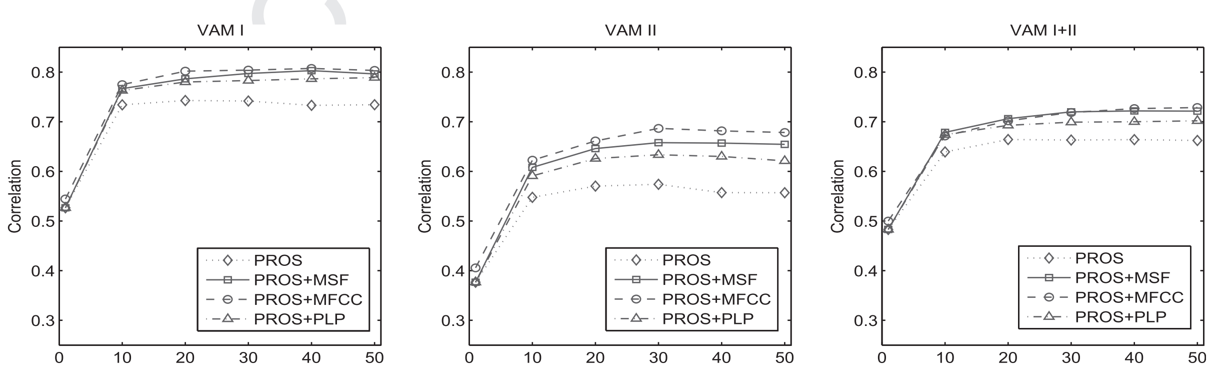

Fig. 16. Mean correlations for prosodic and combined features selected using SFS (VAM).

Table 9
Regression results for continuous emotions on the VAM database using prosodic and combined features.

Dataset	Feature	Correlation (r)			Absolute error (e)			Average	
		Valence	Activation	Dominance	Valence	Activation	Dominance	$\bar{r}$	$\bar{e}$
VAM I	PROS only	0.58	0.85	0.82	0.11	0.16	0.15	0.75	0.14
	PROS + MSF	0.66	0.90	0.86	0.10	0.13	0.13	0.81	0.12
	PROS + MFCC	0.66	0.90	0.86	0.10	0.13	0.13	0.81	0.12
	PROS + PLP	0.62	0.90	0.85	0.11	0.13	0.14	0.79	0.13
	Grimm et al. (2007a)	N/A	N/A	N/A	N/A	N/A	N/A	0.71	0.27
	Human	0.49	0.78	0.68	0.17	0.25	0.20	0.65	0.21
VAM II	PROS only	0.38	0.70	0.64	0.14	0.17	0.15	0.57	0.15
	PROS + MSF	0.48	0.78	0.73	0.14	0.15	0.14	0.66	0.14
	PROS + MFCC	0.54	0.79	0.73	0.13	0.15	0.14	0.69	0.14
	PROS + PLP	0.45	0.75	0.70	0.14	0.16	0.14	0.63	0.15
	Grimm et al. (2007a)	N/A	N/A	N/A	N/A	N/A	N/A	0.43	0.23
	Human	0.48	0.66	0.54	0.11	0.19	0.14	0.56	0.15
VAM I+II	PROS only	0.47	0.77	0.75	0.13	0.17	0.15	0.66	0.15
	PROS + MSF	0.55	0.82	0.80	0.13	0.15	0.14	0.72	0.14
	PROS + MFCC	0.56	0.83	0.80	0.12	0.15	0.14	0.73	0.14
	PROS + PLP	0.52	0.82	0.78	0.13	0.16	0.15	0.71	0.15
	Grimm et al. (2007a)	N/A	N/A	N/A	N/A	N/A	N/A	0.60	0.24
	Human	0.49	0.72	0.61	0.14	0.22	0.17	0.61	0.18

(Grimm et al., 2007b), good estimation results are also achieved for *activation* and *dominance* on VAM I + II, but *valence* is still poorly estimated with 0.46 correlation reported. This corroborates the poor *anger* vs. *joy* classification results in Tables 5 and 6, as the two emotions are with the same *activation* level but opposite *valence*. Since it has also been shown that *anger* vs. *sadness*, emotions with opposite *activation* but similar *valence*, can be separated very well, both the Berlin (discrete and acted) and VAM (continuous and natural) databases show the tendency of the SER systems to recognize *activation* more accurately than *valence*.

Moreover, as might also be noticed from Figs. 15 and 16 as well as Tables 8 and 9, regression performance on VAM I is always superior to the performance on VAM II, regardless of the feature types. This is because VAM I contains utterances from "very good" speakers who are characterized by a high level of activity and a wide variety of emotions, while VAM II consists of utterances from "good" speakers that, though possessing high activity as well, produce a smaller scope of emotions (e.g. *anger* only) (Grimm et al., 2008). VAM I + II, as a combination of VAM I and VAM II, exhibits recognition performance intermediate between VAM I and VAM II.

5.2.2. Cross-database evaluation

In (Grimm et al., 2007a), the authors mapped the continuous-valued estimates of the emotion primitives into discrete emotion categories on the EMA corpus (Lee et al., 2005). In this section, a similar experiment is performed to apply the continuous primitives to classify the discrete emotions in the Berlin database. More specifically, since the regression performance varies for the three primitives as shown in the previous experiment, several binary classification tasks are employed here to further evaluate the primitives separately.

First, three SVRs are trained on the VAM I + II dataset using the combined prosodic and proposed features, one for each of the three primitives. Each SVR uses the set of features selected by SFS on VAM I + II that yield the highest correlation. The Berlin database is then used as a separate source for evaluation, where each file is assigned three predicted primitive values by the three trained regressors. The three estimated primitive values are then used as features for classifying the seven emotions, yielding an overall recognition rate of 52.7% (49.9%) under 10-fold cross-validation with (without) SN. Even though this accuracy is unattractive, it is still better than the 14% random chance, indicating that the continuous primitives do convey useful information about the discrete emotion classes in the Berlin database. The low recognition rate could be due to inadequacy of using the three primitives. Also, no "mutual information" relating the emotion data of the two databases is available. Such information could have been produced by having the same human subjects rate both databases on both the discrete and primitive scales.

Three two-class classification tasks are designed to test the primitive features: (1) *anger* vs. *joy*, (2) *anger* vs. *sadness*, and (3) *anger* vs. *fear*, which mainly involve (though not limited to) the recognition of *valence*, *activation*, and *dominance*, respectively. Ideally, if the three primitives were well recognized, *valence*, *activation*, and *dominance* would be the features providing the best recognition performance for tasks (1), (2), and (3), respectively.

Recognition results are listed in Table 10 where SN has been applied (in the same way as did in Section 5.1). As can be seen from the table, high classification accuracy is achieved when discriminating *anger* vs. *sadness* (100%) and *anger* vs. *fear* (83.7%) using only one primitive feature,

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013

Table 10
Binary classification results with three primitive features (cross-database).

	Valence (%)	Activation (%)	Dominance (%)	All (%)
Anger vs. *joy*	63.6	61.6	61.1	58.6
Anger vs. *sadness*	65.6	100	99.5	99.0
Anger vs. *fear*	64.8	79.6	83.7	81.6

namely *activation* and *dominance*, respectively. Although the unrealistic 100% accuracy may be partly due to the acted emotion in the Berlin database, the result still corroborates previous results that showed the *activation* primitive to be well recognized by machine. Combining all three primitive features, however, does not improve the recognition performance. One notable difficulty, as expected, arises when using *valence* to classify *anger* and *joy*, which are with similar activations and opposite valence, resulting in 63.6% recognition rate only. These cross-database results reinforce our previous findings that *activation* and *valence* are primitives with the best and worst estimation performance, respectively. In addition, the seven-class FDR scores for the three primitive features are *valence*: 0.6 (0.4), *activation*: 4.4 (3.7), and *dominance*: 3.9 (3.3) with (without) SN, which again indicate that the *valence* feature has the lowest discrimination power.

6. Conclusion

This work presents novel MSFs for the recognition of human emotions in speech. An auditory-inspired ST representation is acquired by deploying an auditory filterbank as well as a modulation filterbank, to perform spectral decomposition in the conventional acoustic frequency domain and in the modulation frequency domain, respectively. The proposed features are then extracted from this ST representation by means of spectral measures and linear prediction parameters.

The MSFs are evaluated first on the Berlin database to classify seven discrete emotions. Typical short-term spectral features and prosodic features are extracted to benchmark the proposed features. Simulation results show that the MSFs outperform MFCC and PLP features when each feature type is used solely, in terms of both FDR scores and recognition accuracy. The proposed features are also shown to serve as powerful additions to prosodic features, as substantial improvement in recognition accuracy is achieved once prosodic features are combined with the MSFs, with up to 91.6% overall recognition accuracy attained. In a LOSO cross-validation test, the MSFs give superior performance except in the case when speaker normalization is applied; MFCC combined with prosodic features outperform MSFs combined with prosodic features.

Besides the classic discrete emotion classification, continuous emotion estimation is further investigated in this study. MFCC features are shown to deliver the best results in the regression experiments. However, the apparent gain of MFCC features over MSFs is mainly due to the former's higher correlation with *valence*. For the other two primitives, the performance of the proposed and MFCC features are comparable. PLP features are found to deliver generally inferior outcomes relative to the other two spectral feature types. Combining spectral and prosodic features further improves the regression results. Promising performance is achieved for estimating *activation* and *dominance*, but lower performance is observed with *valence*, a trend also reported in other continuous emotion recognition studies. Moreover, the continuous emotion primitives are applied to classify discrete emotions. Among the three primitive features, *activation* and *dominance* are shown to be useful for classifying discrete emotions, but further investigation is needed for the *valence* parameter to achieve practical performance figures.

Combining the performance results for the Berlin and VAM databases, the new MSFs appear to offer equally competitive performance with respect to prosodic and MFCC features. Over the decades, MFCCs have been well optimized and demonstrated good performance for automatic speech recognition applications. However, features extracting modulation spectral information have recently demonstrated promising performance for various applications in speech processing (e.g. Falk and Chan, 2008, 2010a,). This paper has demonstrated the potential and promise of the MSFs for emotion recognition. With possible refinement in future work, the performance of modulation domain features could be further improved. Hence, further research on the use of MSFs for SER can be beneficial.

Acknowledgements

The authors would like to thank Dr. Michael Grimm for providing the VAM database and the reviewers for helpful improvement of the paper.

References

Abelin, A., Allwood, J., 2000. Cross linguistic interpretation of emotional prosody. In: Proc. ISCA ITRW on Speech and Emotion, pp. 110–113.

Aertsen, A., Johannesma, P., 1980. Spectro-temporal receptive fields of auditory neurons in the grass frog. I. Characterization of tonal and natural stimuli. Biol. Cybernet. 38, 223–234.

Atlas, L., Shamma, S., 2003. Joint acoustic and modulation frequency. EURASIP J. Appl. Signal Process., 668–675.

Barra, R., Montero, J., Macias-Guarasa, J., D'Haro, L., San-Segundo, R., Cordoba, R., 2006. Prosodic and segmental rubrics in emotion identification. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, pp. 1085–1088.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2006. Combining efforts for improving automatic classification of emotional user states. In: Proc. IS-LTC, pp. 240–245.

Bishop, C., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A database of German emotional speech. In: Proc. Interspeech, pp. 1517–1520.

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

Chang, C.-C., Lin, C.-J., 2009. LIBSVM: A library for support vector machines. Tech. rep., Department of Computer Science, National Taiwan University. Software available at: <http://www.csie.ntu.edu.tw/cjlin/libsvm>.

Chih, T., Ru, P., Shamma, S., 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Amer. 118, 887–906.

Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T., 2008. Fear-type emotion recognition for future audio-based surveillance systems. Speech Commun. 50, 487–503.

Cowie, R., Douglas-Cowie, E., 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: Proc. Internat. Conf. on Spoken Language Processing, Vol. 3, pp. 1989–1992.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human–computer interaction. IEEE Signal Process. Mag. 18, 32–80.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Audio Speech Language Process. 28, 357–366.

Depireux, D., Simon, J., Klein, D., Shamma, S., 2001. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J. Neurophysiol. 85, 1220–1234.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: Towards a new generation of databases. Speech Commun. 40, 33–60.

Ekman, P., 1999. Basic Emotions. John Wiley, New York, pp. 45–60.

Ewert, S., Dau, T., 2000. Characterizing frequency selectivity for envelope fluctuations. J. Acoust. Soc. Amer. 108, 1181–1196.

Falk, T.H., Chan, W.-Y., 2008. A non-intrusive quality measure of dereverberated speech. In: Proc. Internat. Workshop for Acoustic Echo and Noise Control.

Falk, T.H., Chan, W.-Y., 2010a. Modulation spectral features for robust far-field speaker identification. IEEE Trans. Audio Speech Language Process. 18, 90–100.

Falk, T.H., Chan, W.-Y., 2010b. Temporal dynamics for blind measurement of room acoustical parameters. IEEE Trans. Instrum. Meas. 59, 978–989.

Fletcher, H., 1940. Auditory patterns. Rev. Modern Phys. 12, 47–65.

Giannakopoulos, T., Pikrakis, A., Theodoridis, S., 2009. A dimensional approach to emotion recognition of speech from movies. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, pp. 65–68.

Glasberg, B., Moore, B., 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Res. 47, 103–138.

Grimm, M., Kroschel, K., Mower, E., Narayanan, S., 2007a. Primitives-based evaluation and estimation of emotions in speech. Speech Commun. 49, 787–800.

Grimm, M., Kroschel, K., Narayanan, S., 2007b. Support vector regression for automatic recognition of spontaneous emotions in speech. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Vol. 4, pp. 1085–1088.

Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera am Mittag German audio-visual emotional speech database. In: Proc. Internat. Conf. on Multimedia & Expo, pp. 865–868.

Gunes, H., Piccard, M., 2007. Bi-modal emotion recognition from expressive face and body gestures. J. Network Comput. Appl. 30, 1334–1345.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87, 1738–1752.

Hsu, C.-C., Chang, C.-C., Lin, C.-J., 2007. A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University.

Intl. Telecom. Union, 1993. Objective measurement of active speech level. ITU-T P.56, Switzerland.

Intl. Telecom. Union, 1996. A silence compression scheme for G.729 optimized for terminals conforming to ITU-T recommendation V.70. ITU-T G.729 Annex B, Switzerland.

Ishi, C., Ishiguro, H., Hagita, N., 2010. Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech. EURASIP J. Audio Speech Music Process., article ID 528193, 12 pages.

Kaiser, J., 1990. On a simple algorithm to calculate the 'energy' of a signal. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Vol. 1, pp. 381–384.

Kanederaa, N., Araib, T., Hermanskyc, H., Pavel, M., 1999. On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Commun. 28, 43–55.

Kittler, J., 1978. Feature set search algorithms. Pattern Recognition Signal Process., 41–60.

Lee, S., Yildirim, S., Kazemzadeh, A., Narayanan, S., 2005. An articulatory study of emotional speech production. In: Proc. Interspeech, pp. 497–500.

Lugger, M., Yang, B., 2008. Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Vol. 4, pp. 4945–4948.

Morgan, N., Zhu, Q., Stolcke, A., Sonmez, K., Sivadas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellis, D., Doddington, G., Chen, B., Cetin, O., Bourlard, H., Athineos, M., 2005. Pushing the envelope – aside. IEEE Signal Process. Mag. 22, 81–88.

Mozziconacci, S., 2002. Prosody and emotions. In: Proc. Speech Prosody, pp. 1–9.

Nwe, T., Foo, S., De Silva, L., 2003. Speech emotion recognition using hidden markov models. Speech Commun. 41, 603–623.

Picard, R., 1997. Affective Computing. The MIT Press.

Rabiner, L., Juang, B., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ.

Scherer, K., 2003. Vocal communication of emotion: A review of research paradigms. Speech Commun. 40, 227–256.

Scherer, S., Schwenker, F., Palm, G., 2007. Classifier fusion for emotion recognition from speech. In: Proc. IET Internat. Conf. on Intelligent Environments, pp. 152–155.

Schuller, B., Seppi, D., Batliner, A., Maier, A., Steidl, S., 2006. Emotion recognition in the noise applying large acoustic feature sets. In: Proc. Speech Prosody.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2007a. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In: Proc. Interspeech, pp. 2253–2256.

Schuller, B., Seppi, D., Batliner, A., Maier, A., Steidl, S., 2007b. Towards more reality in the recognition of emotional speech. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, Vol. 4. pp. 941–944.

Shami, M., Verhelst, W., 2007. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun. 49, 201–212.

Shamma, S., 2001. On the role of space and time in auditory processing. Trends Cogn. Sci. 5, 340–348.

Shamma, S., 2003. Encoding sound timbre in the auditory system. IETE J. Res. 49, 193–205.

Slaney, M., 1993. An efficient implementation of the Patterson–Holdsworth auditory filterbank. Tech. rep., Apple Computer, Perception Group.

Sun, R., E., M., Torres, J., 2009. Investigating glottal parameters for differentiating emotional categories with similar prosodics. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing, pp. 4509–4512.

Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). Elsevier Science, Speech Coding and Synthesis (Chapter 14).

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: Resources, features, and methods. Speech Commun. 48, 1162–1181.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007. Combining frame and turn-level information for robust recognition of emotions within speech. In: Proc. Interspeech, pp. 2225–2228.

Wollmer, 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: Proc. Interspeech, pp. 597–600.

Wu, S., Falk, T., Chan, W.-Y., 2009. Automatic recognition of speech emotion using long-term spectro-temporal features. In: Proc. Internat. Conf. on Digital Signal Processing, pp. 1–6.

Zhou, G., Hansen, J., Kaiser, J., 2001. Nonlinear feature based classification of speech under stress. IEEE Trans. Audio Speech Language Process. 9, 201–216.

Please cite this article in press as: Wu, S. et al., Automatic speech emotion recognition using modulation spectral features, Speech Comm. (2010), doi:10.1016/j.specom.2010.08.013