SALIENCE BASED LEXICAL FEATURES FOR EMOTION RECOGNITION

Kalani Wataraka Gamage[1,2], Vidhyasaharan Sethu[1], Eliathamby Ambikairajah[1,2]

[1] School of Electrical Engineering and Telecommunications, UNSW, Australia
[2] ATP Research Laboratory, DATA61, CSIRO, Australia

ABSTRACT

In this paper we focus on the usefulness of verbal events for speech based emotion recognition. In particular, the use of phoneme sequences to encode verbal cues related to the expression of emotions is proposed and lexical features based on these phoneme sequences are introduced for use in automatic emotion recognition systems where manual transcripts are not available. Secondly, a novel estimate of emotional salience of verbal cues, applicable to both phoneme sequences and words, is presented. Experimental results on the IEMOCAP database show that the proposed automatic phoneme sequence based features can achieve an Unweighted Average Recall (UAR) of 49% with proposed salience measure. Further, the proposed salience measure can lead to an UAR of 64% when using manual word transcriptions. Both of these are the highest UARs reported on the IEMOCAP database for systems using lexical features extracted from automatic and manual transcripts respectively.

Index Terms— speech based emotion recognition, human-computer interactions, lexical features, emotional salience

1. INTRODUCTION

Being the primary modality of natural human communication, speech carries both direct and indirect cues about human emotion. Consequently speech based emotion recognition is expected to play a key role in emotion-sensitive technologies in many fields including human-computer interfacing, entertainment, learning, etc. [1]. Current speech based emotion recognition systems typically employ acoustic features such as pitch, energy, voice quality features and cepstral features as the basis for recognition since they reflect the style of human speech expression and other types of features are less commonly used [2-4].

However, a number of studies have suggested that the combination of these acoustic features with lexical features, either at feature level or decision level, can lead to more accurate emotion recognition systems [5-11]. Traditionally lexical features are generated from manual transcripts [10, 11] and in practical emotion recognition setting, it has been suggested that transcripts based on automatic speech recognition (ASR) be used in place of manual transcripts [12, 13]. However, ASR systems generally do not provide all the lexical information that can be expected to be useful in emotion recognition. For e.g., most ASR systems do not indicate variations in pronunciations or label emotionally salient non-speech events such as laughter and other verbal 'gestures'.

In this paper, we propose the use of phoneme sequences as indicators of different verbal expressions that can capture both spoken content and verbal gestures that do not always depend on underlying language and content. While individual phonemes may not be important in terms of indicating verbal events of interest, a phoneme sequence can be expected to be indicative of both spoken content and verbal gestures.

It has been recognised that non-linguistic vocalizations can provide additional information toward emotion recognition [13]. Previous attempts to capture them include modelling a few of the non-vocal cues such as silences, laughs and sighs as separate models in order to recognise them from speech. As pointed out in [13], this requires training data with sufficient number of occurrences (of sufficient duration) of these non-verbal cues in order for the models to be trained. This can lead to only a small number of nonverbal cues being useful for recognition task as in [13]. The proposed approach of using phoneme sequences to capture verbal gestures is potentially capable of avoiding most of these difficulties as no specific modelling is involved. Finally the most salient gesture representations can be emphasised through appropriate weighing.

The advantage of explicitly capturing information about these types of verbal events (herein we use the term 'verbal events' to refer to both spoken content and verbal gestures) is that they can represent a great deal of information relevant to emotion recognition. However, relying on purely acoustic features to capture similar information can be problematic since the information may be spread over long and variable durations and acoustic features are typically estimated from frames of fixed duration.

Bag-Of-Words (BOW) [11], Bag-of-n-grams [9, 12] and their refinements such as log term frequency (log TF) [13], adopted from text classification are the most commonly used lexical features in speech emotion recognition. As previously mentioned these lexical features are extracted from either manual or ASR transcripts and do not capture verbal gestures unrelated to spoken content. Furthermore, these are high dimensional vectors since their dimensionality is proportional to total number of words or n-grams (in the case of Bag-of-n-grams) present in the training data. Different approaches such as stop word removal, potter stemming or data driven methods such as salience or information gain based dimensionality reduction methods are therefore used to reduce the vocabulary prior to vector space modelling [2, 12]. Despite this, the final dimensionality is typically still very high and most elements will be zeros for any given utterance. Some other approaches to representing lexical information include the use of belief networks for key word/ phrase spotting [6, 7]; and the use of string kernel mapping on data without explicit calculation of high dimensional features [12].

Recently, a new lexical representation called the lex-eVector has been proposed, wherein, each word is first given a weight to represent its emotional salience for each emotion class and then, each utterance's emotional salience toward each emotion is derived as the mean of emotional saliences of the words in the given utterance [11]. Emotiorn recognition systems using the low dimensional lex-eVector have been shown to outperform those

using higher dimensional Bag-Of-Words feature vectors. The concept of 'emotional salience of words' was initially used in the context of emotion recognition in [8], which introduced self-mutual information to measure emotional salience. In this paper, we introduce a new weighting scheme for use in a lex-eVector representation that can capture the emotional salience of both spoken content and verbal gestures when using either word level transcripts (spoken content) or phoneme sequences (spoken content and verbal gestures).

2. PROPOSED PHONEME SEQUENCES BASED FEATURES

We propose the use of phoneme sequences to uniquely represent or encode specific sounds to represent spoken content and verbal gestures. Although single phonemes may not capture any specific significant information, phoneme sequences of different lengths can be more specific and therefore identify specific verbal events.

The main idea is to use a phoneme recognizer to decode a given utterance into a stream of phonemes and then identify all possible phoneme sequences of a predetermined length, p. Following this, a vocabulary or a phoneme sequence dictionary is developed comprising of all distinct phoneme sequences of length P present in the training data.

Each utterance, u, can then be represented as a bag-of-phoneme sequences (BOP) vector, $b_p(u)$, of length equal to the size of phoneme sequence dictionary and then indicate the presence of each phoneme sequence within that utterance as follows:

$$b_p(u) = [c_1, c_2, \dots c_M]^T \quad (1)$$

Where, M is the number of distinct phoneme sequences of length p present in the training data and c_i denotes the number of times the i^{th} phoneme sequence occurs in utterance u.

The bag-of-phoneme sequences (BOP) are analogous to bag-of-words (BOW) representation and are intended to be used in lieu of bag-of-words. Finally, these bag-of-phoneme sequences vectors are mapped to lower dimensional weighted lexical feature by applying a suitable weighting scheme as outlined in section 3.

The use of phoneme sequences instead of words potentially offer the advantage of being able to capture verbal gestures (non-speech) as well as spoken content. Also, relying on sub-word lexical units may make the system less sensitive to changes in content or amount of training data and as suggested in [2].

3. LEXICAL FEATURE EXTRACTION

3.1. Weighted Lexical Features Approach

We represent the weighted lexical features as a low dimensional vector (v_e) indicating a given utterance's inclination towards each of the emotional classes of interest based on the concept of emotional salience [8] and lex-eVectors [11]. The idea behind this approach is the weighting of each lexical unit (word or phoneme sequences) by a value that is indicative of the relevance of that lexical unit to predicting the emotions of interest. The weighted lexical feature vector corresponding to an utterance is given by:

$$v_e(u) = \frac{1}{K}\mathbf{\Phi} b_p(u) \quad (2)$$

where, $v_e(u)$ is the $n\times1$ weighted lexical feature vector; K is the total number of lexical units in utterance u; $\mathbf{\Phi}$ is an $n\times M$ matrix whose elements $\phi_{j,k}$ are the weights corresponding to the j^{th} emotion for the k^{th} lexical unit; n is the total number of emotions of interest; M is the size of lexical dictionary and $b_p(u)$ can be either the BOP vector estimated as per (1) or the traditional BOW vector.

3.2. Proposed Relative Frequency based Lexical Feature (LRF)

In practical situations and in most databases comprising of elicited real life emotions (not acted), the emotions are not always expressed as full blown emotions. Utterances will also not be fully composed of emotionally salient words but comprise of a mixture of emotionally salient and non-salient words. The relative frequency based lexical feature is proposed to take into account the emotional salience of lexical units in terms relative frequency of occurrence in the training data across utterances corresponding to the emotions of interest. In addition, a lexical unit found in utterances corresponding to more than one emotion at low frequencies of occurrences should not lead to penalties. The proposed lexical unit weighting is defined as the relative frequency of the k^{th} lexical unit within the j^{th} emotion class with respect to a normalized frequency of that lexical unit's occurrence within other emotions. Therefore this weight is relaxed and shows fewer penalties compared to the weighting in [11]. The proposed weighting is given by:

$$\phi_{j,k} = \frac{\eta_j(w_k)}{1 + \frac{\hat{\eta}_j(w_k)}{n-1}} \quad (3)$$

where, $\eta_j(w_k)$ and $\hat{\eta}_j(w_k)$ denote the number occurrences of the k^{th} lexical unit (word or phoneme sequence), w_k, in utterances corresponding to the j^{th} emotion and the number of occurrences of w_k in utterances not corresponding to the j^{th} emotion respectively; and n is the total number of emotions of interest.

The relative frequency based weighted lexical features, (denoted as LRF), is then $v_e(u)$ given by equations (1), (2) and (3).

We also introduce a further modification to the proposed LRF by incorporating the maximum weights across the lexical units of an utterance in addition to the average weights. The modified LRF (denoted as mLRF) is then given by:

$$\overline{v_e}(u) = [v_e(u)^T \; m_e(u)^T]^T \quad (4)$$

where, $\overline{v_e}(u)$ is the $2n\times1$ mLRF vector; $v_e(u)$ is the $n\times1$ LRF vector and $m_e(u)$ is the $n\times1$ vector given as follows:

$$m_e(u) = [\, a_1(u),\, a_2(u),\, \dots,\, a_n(u)] \quad (5)$$

where, n is the number of emotions of interest and $a_j(u)$ corresponding to the j^{th} emotion is given as

$$a_j(u) = \max_i \phi_{j,i} \quad (6)$$

where, i denotes the indices corresponding to non-zero elements of $b_p(u)$.

The inclusion of $m_e(u)$ in the modified LRF is done to avoid watering down of the weights in long utterances with a small number of emotionally salient words. It should be noted that both the

proposed LRF and mLRF vectors can be applied to bag-of-phoneme sequences (BOP) or bag-of-words (BOW) representations and both combinations are evaluated.

4. EXPERIMENTS

4.1. Dataset Description

The IEMOCAP database [15] provided by the University of Southern California (USC) is used for the experiments in this paper. The database is collected in the form of interactive dyadic sessions between actors on both scripted and improvised scenarios. The database consists of 12 hours of data from 10 speakers, each session recorded between one male and one female actor. Speech dialogues are segmented to utterances and are annotated by 3 annotators. We considered emotion classes: neutral, angry, sad and happy where excitement class is merged into happy class, to balance data distribution between classes. As with many experiments on this database, we also used utterances for which majority agreement on emotion labels between annotators are present. The numbers of speech utterances available per emotional class are listed in Table 1. Word, syllable and phoneme level transcripts obtained via force alignment of manual word level transcriptions with the speech signals are provided for each utterance [15].

Table 1. Number of utterances per emotion class

	Neutral	Angry	Sad	Happy
Complete set	1708	1103	1084	1636
Improvised only	1099	289	608	946

4.2. Experiment Objectives and Baseline Systems

The purpose of the experiments reported in this paper are to demonstrate that the proposed LRF and mLRF features (applied to both BOP and BOW) are effective in representing emotionally salient verbal events. The specific experiments are listed below:

Using Transcripts: The proposed LRF and mLRF features, extracted from manual word transcripts, are compared with established BOW and lex-e-vector features [11] in order to determine if the proposed lexical features are superior to BOW and lex-e-vector features.

Without Transcripts: The LRF and mLRF features, extracted from BOP representations obtained using a phone decoder, are compared to established lexical features extracted from word transcripts obtained using an ASR system [10]. This comparison is used to evaluate the performance of the proposed features in the more common scenario where manual transcripts are not available.

Fusion with Acoustic Features: The impact of fusing lexical features with acoustic features is also investigated. Here the 2009 Interspeech feature set (ISO9) extracted using the openSMILE toolkit [17] is used as the acoustic feature set.

Scripted Vs Improvised Speech: The IEMOCAP dataset contains both scripted and improvised speech from each speaker and the scripted speech content is identical across all speakers of the same gender [15]. Consequently, the speech content in training and test sets will be identical for the scripted data and therefore experiments were conducted both with and without the scripted speech segments.

Results from the literature based on experiments conducted using same database with identical training and test partitions and cross validation methods are used as baseline systems [10, 11] in section 4.4.

4.3. Experiment settings

Classification experiments were conducted in a leave-one-speaker-out cross validation manner with a Random Forest (RF) classifier that uses 100 trees. Additionally, in order to facilitate direct comparisons with [11] and [10], the experiments were repeated using a linear SVM classifier as the back-end and these results are reported within parentheses in section 4.4 (Table 2 and Table 3). The word level transcripts provided with the database were used to generate all word based features, and a phoneme recogniser developed by Brno University of Technology for English, with a reported Phoneme Error Rate (PER) of 24.4% on TIMIT [16], was used to generate phoneme based features. Phoneme sequences of lengths 2 and 3 are considered in our experiments and denoted as 'seq2' and 'seq3'.

4.4. Results and Discussion

4.4.1. Using Transcripts

Table 2 reports the performances of the proposed LRF and mLRF feature representations based on manual word transcriptions and compares them to established lex-e-vector and BOW features [11]. From these results, the proposed LRF and mLRF features clearly outperform the established lexical features. From Table 2, it is also clear that the modified LRF (mLRF) feature has a slight advantage over the LRF features. These results indicate that the proposed emotional salience weighting technique is superior to the weighting approach used in [11] and the BOW features [11] which do not have any emotional salience weighting. Both the LRF and the mLRF weighing schemes do not specifically penalise words that appear in speech corresponding to multiple emotions and only the relative number of occurrences within each emotional class matters, since words semantically connected to one emotion can also occasionally occur within another emotional class. For example, the word "hell" is semantically related to anger but can also be used in a joke. Additionally, the mLRF features are less affected by the presence of a larger number of neutral words (low emotional salience) compared to both LRF features and other existing lexical features due to the inclusion of maximum weights in addition to average weights (refer to equations 4-6).

Table 2. Weighted and Unweighted average recall (WAR and UAR) evaluated on complete IEMOCAP dataset using manual transcripts. UAR and WAR obtained using linear SVM back-end are reported in parentheses.

Feature set	**UAR**	**WAR**
Lex eVector [11] (Linear SVM)	Not reported	57.4
BOW [11] (Linear SVM)	Not reported	56
LEX-T [10] (RBF SVM)	55.3	Not reported
Proposed LRF	62.0 (59.9)	61.9 (60.2)
Proposed mLRF	**64.0 (60.5)**	63.8 (60.7)

4.4.2. Without Transcripts

Table 3 reports the performances of the proposed LRF and mLRF feature representations obtained using phoneme sequences. These are compared with ASR based lexical features that were previously proposed in [10]. As indicated in Table 3, the proposed phoneme sequence based LRF and mLRF features have led to 48% and 49.9% UAR respectively compared to 43.8% for ASR based features [10].

Table 4 shows the phoneme sequences corresponding to the highest LRF weights ($\phi_{j,k}$ in eqn. 3) and provides some insight to the underlying behaviour. Because the phoneme recogniser converts speech into a stream of phonemes irrespective of language, we can see that interjections (words/expressions that are distinct from the rest of the sentence and generally signify emotions or spontaneous feelings) as well as sounds which resonate with emotions are highlighted here, such as 'eh-aa-hh' (expression of disgust), 'iy-ae-ay' (cheering) and 'm-m-m' (sound of long audible breaths or sighs).

Table 3. Weighted and Unweighted average recall (WAR and UAR) evaluated on complete IEMOCAP dataset without using transcripts. UAR and WAR obtained using linear SVM back-end are reported in parentheses.

Feature set	**UAR**	**WAR**
LEX-ASR [10] (RBF SVM)	43.8	Not reported
Proposed LRF (seq2)	48.0 (46.7)	46.9 (46.7)
Proposed LRF (seq3)	47.3 (46.3)	45.7 (45.8)
Proposed mLRF (seq2)	**49.9 (47.8)**	49.0 (47.8)
Proposed mLRF (seq3)	47.7 (47.8)	46.1 (47.1)

Table 4. Examples of phoneme sequence of length three (Seq3) corresponding to highest mLRF weights in emotional classes: angry, happy and sad

Angry	'aa-hh-hh', 'sh-iy-iy', 'eh-aa-hh', 'iy-iy-n', 'iy-jh-iy', 'ay-ay-t'
Happy	'ow-hh-iy', 'ay-ay-hh', 'hh-aw-aa', 'iy-ae-ay', 'hh-ay-eh'
Sad	'm-m-m', 'n-n-m', 'pau-m-m', 'm-r-hh', 'pau-n-n', 'm-t-ah'

4.4.3. Fusion with Acoustic Features

The performances of systems that fused the proposed lexical features (mLRF) with established acoustic features are reported in Table 5. As expected the fusion of complementary acoustic and lexical features led to higher recognition rates than using only acoustic or lexical features. In particular, the fusion of mLRF based on manual transcripts with the IS09 acoustic features resulted in an UAR of 67.3%. Finally, it should be noted that the BUT phoneme recognizer used in these experiments has a reported error rate of 24.2% [16] but the proposed features are still able to use the automatic phoneme transcripts to capture lexical information.

4.4.4. Scripted and Improvised Vs Improvised only

As previously mentioned, the proposed lexical features were also evaluated on the improvised speech subset of the IEMOCAP database and the performances obtained are reported in Table 6. It should be noted that when using only improvised speech, there is no overlap in speech content between the training and test sets. It is interesting to observe that under these conditions, the performance of the system based on manual transcripts dropped by more than 10% (UAR) while the performance of phoneme sequence based lexical features dropped by only 3%. This observation supports the idea that features based on phoneme sequences are more suitable than higher level lexical features for systems that are trained on small datasets.

Table 5. Weighted and Unweighted average recall (WAR and UAR) for complete IEMOCAP dataset

Feature set	**UAR**	**WAR**
IS09 (Acoustic features)	57.2	56.9
mLRF + IS09 (using transcripts)	67.3	67.2
mLRF (seq2) + IS09 (without transcripts)	58.2	57.4
mLRF (seq3) + IS09 (without transcripts)	*59.2*	*58.6*

Table 6. UAR of systems trained and evaluated on improvised speech only from IEMOCAP database

Feature set	**UAR**	**WAR**
mLRF (using transcripts)	52.2	57.3
mLRF (seq2) (without transcripts)	47.5	52.3
mLRF (seq3) (without transcripts)	45.3	50.5

5. CONCLUSIONS

This paper has presented novel lexical features which have the ability to capture emotionally salient verbal gestures that can be extracted from different types of verbal events including words and phoneme sequences. In particular, the proposed LRF and mLRF features are able to incorporate an estimate of emotional salience better than the previously introduced lex-e-vector and are more suited to emotion recognition than the standard bag-of-words features. The LRF and mLRF features can also be extracted from phoneme sequences which can encode simple verbal gestures including sounds that will not be captured by ASR but still carry information relevant to emotion recognition.

6. REFERENCES

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

[2] B. Schuller and A. Batliner, Computational *paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[3] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern *Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] V. Sethu, J. Epps, and E. Ambikairajah, “Speech based emotion recognition,” in *Speech and Audio Processing for Coding, Enhancement and Recognition*. Springer, 2015, pp. 197–228.

[5] J. Liscombe, G. Riccardi and D. Hakkani-Tur, "Using Context to Improve Emotion Detection in Spoken Dialog Systems", *Academiccommons.columbia.edu*, 2005. [Online]. Available:http://academiccommons.columbia.edu/catalog/ac:161847. [Accessed: 23- Sep- 2015].

[6] B. Schuller, G. Rigoll and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[7] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang and G. Rigoll, "Speaker Independent Speech Emotion Recognition by Ensemble Classification", *2005 IEEE International Conference on Multimedia and Expo*.

[8] Chul Min Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs", *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293-303, 2005.

[9] B. Schuller, F. Metze, S. Steidl, A. Batliner, F. Eyben and T. Polzehl, "Late fusion of individual engines for improved recognition of negative emotion in speech - learning vs. democratic vote",*2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[10] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, A. Vembu and R. Prasad, “Emotion Recognition using Acoustic and Lexical Features”, in *INTERSPEECH 2012*, 2012.

[11] Q. Jin, C. Li, S. Chen and H. Wu, "Speech emotion recognition with acoustic and lexical features", *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[12] B. Schuller, A. Batliner, S. Steidl and D. Seppi, "Emotion recognition from speech: Putting ASR in the loop", *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[13] B. Schuller, R. Mueller, F. Eyben, J. Gast, B. Hoernler, M. Woellmer, G. Rigoll, A. Hoethker and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application", *ELSEVIER SCIENCE BV*, 2009.

[14] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller and S. Steidl, "Emotion Recognition using Imperfect Speech Recognition", in *INTERSPEECH 2010*, pp. 478-481.

[15] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee and S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database", *Lang Resources & Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

[16] P. Schwarz, Phoneme recognition based on long temporal context, Ph.D. thesis, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.

[17] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the international conference on Multimedia.* ACM, 2010, pp. 1459–1462.