Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 87-98

A Novel Trajectory-based Spatial-Temporal Spectral Features for Speech Emotion Recognition

Chun-Min Chang*, Wei-Cheng Lin* and Chi-Chun Lee*

Abstract

Speech is one of the most natural form of human communication. Recognizing emotion from speech continues to be an important research venue to advance human-machine interface design and human behavior understanding. In this work, we propose a novel set of features, termed trajectory-based spatial-temporal spectral features, to recognize emotions from speech. The core idea centers on deriving descriptors both spatially and temporally on speech spectrograms over a sub-utterance frame (e.g., 250ms) - an inspiration from dense trajectory-based video descriptors. We conduct categorical and dimensional emotion recognition experiments and compare our proposed features to both the well-established set of prosodic and spectral features and the state-of-the-art exhaustive feature extraction. Our experiment demonstrate that our features by itself achieves comparable accuracies in the 4-class emotion recognition and valence detection task, and it obtains a significant improvement in the activation detection. We additionally show that there exists complementary information in our proposed features to the existing acoustic features set, which can be used to obtain an improved emotion recognition accuracy.

Keywords: Emotion Recognition, Speech Processing, Spatial-Temporal Descriptors, Mel-Filter Bank Energy

1. Introduction

The blooming of research effort in affective computing (Picard, 1997) in the past decade has started to enable machines to become capable toward sensing and synthesizing emotional expressive behaviors. Numerous technological applications, e.g., advanced human-machine interface (Bach-y Rita & Kercel, 2003; Swartout *et al.*, 2006) and interactive robotic design(Hollinger *et al.*, 2006; Hogan, Krebs, Sharon & Charnnarong, 1995), and even

* Department of Electrical Engineering, National Tsing Hua University
E-mail: cmchang@gapp.nthu.edu.tw; winston810719@gmail.com; cclee@ee.nthu.edu.tw

emerging cross-cutting research fields, e.g. social signal processing (Vinciarelli, Pantic & Bourlard, 2009) and behavioral signal processing (Narayanan & Georgiou, 2013), have all benefited from the vast amount of research advancements in affective computing. Speech is the most natural form of human communication that encodes both linguistic content and paralinguistic information (Schuller *et al*., 2013), e.g., emotion (Nwe, Foo & De Silva, 2003; Scherer, 2003), gender (Childers & Wu, 1991), age (Dobry, Hecht, Avigal & Zigel, 2011), personality (Mairesse & Walker, 2006), etc. Development of suitable algorithms to robustly model emotional content in speech continues to be a prevalent topic in emotion recognition research.

There exists a vast amount of research in modeling speech acoustics for emotion recognition, topics ranging from lowlevel feature engineering, machine learning algorithms, to even joint feature-label representations (Calvo, D'Mello, Gratch & Kappas, 2014; Lee & Narayanan, 2005; Mower, Matarić & Narayanan, 2011). In this work, we aim at proposing a new set of long-term low-level features, named trajectory-based spatial-temporal spectral features, derived directly from the speech spectrograms to perform emotion recognition. Most of the current speech-based emotion recognition rely on extracting a set of commonly-used short-durational features (acoustic low-level descriptors - LLDs), e.g., those could be related spectral features (e.g., MFCCs), prosodic characteristics (e.g., pitch intonation), voicing quality (e.g., jitter), Teager-energy operater etc (Schuller *et al*., 2007). Then, depending on the choice of emotion recognition framework, researcher would either apply global statistical functionals to be used in statics discriminative framework (e.g., support vector machine (Campbell, Sturim & Reynolds, 2006) or deep neural network (Kim, Lee & Provost, 2013) or using time-series model on these short durational low-level descriptors (e.g., hidden Markov model (Nwe *et al*., 2003; Li *et al*., 2013) in order to incorporate the feature's temporal characteristics to perform utterance-level emotion recognition.

Our proposed features are inherently different with the underlying inspiration coming from the dense trajectory-based video descriptors extraction approach (Wang, Kläser, Schmid & Liu, 2013). Dense trajectory video descriptors are extracted by first densely tracking important points on images over a frame (usually 0.5 - 1s) to forms a set of trajectories. The spatial-temporal descriptors for each trajectory can then be computed to obtain the final set of features. By modeling both the trajectory's temporal course and spatial changes over time, these descriptors have been shown to obtain superior improvement in tasks such as event (Oneata, Verbeek & Schmid, 2013) and motion (Wang, Kläser, Schmid & Liu, 2011) recognition than other key-points based image feature extraction. Our core concept, hence, centers around treating an audio file essentially as a sequence of spectrograms. Then, we compute a suite of spatial-temporal descriptors for each trajectory, i.e., a trajectory refers to a spectral energy profile across a time-frame (i.e., 250ms) of a Mel-filter bank output (MFB). In this work, we utilize these descriptors, i.e., trajectory-based spatial-temporal spectral features, to perform speech-based emotion recognition.

To the best of our knowledge, vast majority of the works in the speech emotion recognition literature have utilized short durational (25ms) LLDs which do not share the same concept with our proposed features. There are a few works that utilized auditory perception-inspired modulation spectral features (Chaspari, Dimitriadis & Maragos, 2014; Chi, Yeh & Hsu, 2012), i.e., temporal characteristic of spectral energy, for emotion recognition; these modulation spectrum features have been demonstrated to be robust under noisy conditions compared to features such as MFCCs and fundamental frequencies. In this work, we perform utterance-level categorical (4-emotion classes) and dimensional (valence and activation) emotion recognition on the USC IEMOCAP database (Lee, Mower, Busso, Lee & Narayanan, 2011). We additionally construct two set of features to compare our trajectory-based spatial-temporal spectral features (Traj-ST) to:

- *Conv-PS*: applying statistical functionals over a frame of conventional acoustic feature set
- *OpEmo-Utt*: state-of-the-art exhaustive utterance-level feature extraction using the OpenSmile toolbox (Eyben, Wöllmer & Schuller, 2010)

Our proposed features obtain comparable unweighted average recall on the task of 4-class emotion recognition and significantly outperform on the task of activation recognition compared to *Conv-PS* and *OpEmo-Utt*. Furthermore, by fusing *Traj-ST* with either *Conv-PS* and/or *OpEmo-Utt*, we achieve an improved recognition rate for the 4-class emotion recognition. It demonstrates the complementary information that our proposed features possess when combining with the well established acoustic features for emotion recognition. The rest of the paper is organized as follows: section 2 describes the database and the trajectory-based spatial-temporal spectral features, section 3 describes experimental setups and results, and section 4 is the conclusion and future work.

2. Research Methodology

2.1 The USC IEMOCAP Database

We utilize a well-known emotion database, the USC IEMOCAP database (Busso *et al.*, 2008), for this work. The database consists of 10 actors grouping in pairs to engage in dyadic face-to-face interactions. The design of the dyadic interactions is meant to elicit natural multimodal emotional displays from the actors. The utterances are annotated with both categorical emotion labels (e.g., angry, happy, sad, neural, etc) and dimensional representations (e.g., valence, activation, and dominance) on the scale of 1 to 5. The categorical labels per utterance are annotated by at least 3 raters, and the dimensional attributes are annotated by at least 2 raters. Given the spontaneous nature of this database and the inter-evaluator agreement is about 0.4, this database remains to be a challenging emotion database for algorithmic advancement. In this work, we conduct two different emotion recognition tasks on this database: 1) four-class emotion recognition 2) three-levels of valence and activation dimension recognition. For the categorical emotion recognitions, the four emotion classes are happy, sad, neutral and angry, and we consider samples with the label of ‘excited’ to be the same as ‘happy’. The labels are determined based on the majority vote. The three levels of valence and activation are defined as: low (0 - 1:67), mid (1:67-3:33), and high (3:33-5), where the value of each sample is computed based on the average of the raters. The following lists the number of samples for each type of labels,

- **Four-Emotion Classes:**
  happy: 531, sad: 576, neutral: 411, angry: 378
- **Arousal Dimension:**
  low: 331, mid: 1228, high: 337
- **Valence Dimension:**
  low 653, mid: 820, high:423

2.2 Trajectory-based Spatial-Temporal Spectral Features

Figure 1 depicts the complete flow of our trajectory-based spatial-temporal spectral features extraction approach. Given an audio file, the following is the steps of the feature extraction:

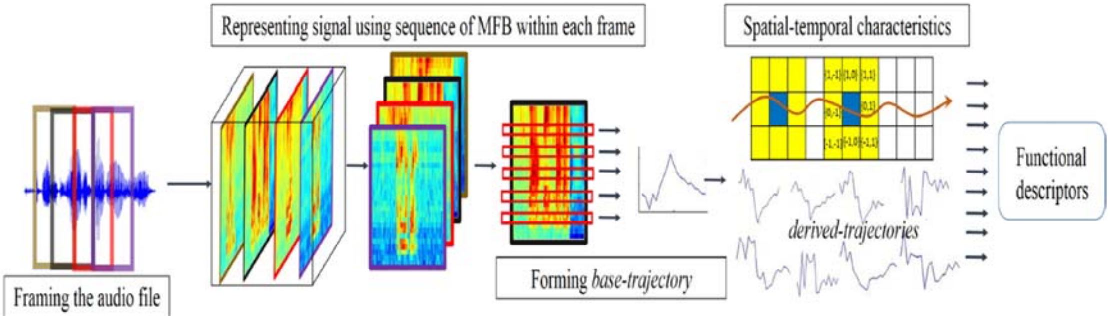


Figure 1. It demonstrates the complete flow of trajectory-based spatial-temporal spectral feature extraction: framing the utterances, representing the signal within each frame using a sequence of MFB, forming base-trajectory of each MFB coefficient, computing grid-based spatial-temporal characteristics and derive 8 additional derived-trajectory, finally frame-level features are extracted by computing 4 statistical functionals on these 9 X 26 trajectories.

(1) Framing the signal:

Segment the entire utterance into regions of frames, where each frame is of length L (L = 250ms , 150ms).There is a 50% overlap between frames.

(2) Representing the segment:

Represent the signal within each frame using a sequence of 26 Mel-filter bank energy (MFB) output - can also be imaged as spectrogram. The window size for MFB is set to be 25ms with 50% overlap. The upper bound of frequency for MFB computation is capped at 3000 Hz.

(3) Forming base-trajectory:

The energy profile for each of the 26 filter output form a base-trajectory over the duration of each frame.

(4) Computing spatial-temporal characteristics:

For each base-$trajectory_i$, at t = 1, we compute the first-order difference with respect to its neighboring grid (8 total: marked as yellow in Figure 1); then we move along the time axis and compute these grid differences until we reach the end of frame. Hence, we obtain 8 extra trajectories (so called, derived-trajectories) to form a total of 9 trajectories (1 base-trajectory+8 derived-trajectories) per frame for each of the 26 filter outputs (a real example of trajectories can be seen in Figure 1).

(5) Frame-level spatial-temporal descriptors:

We derive the final frame-level trajectory-based spatial-temporal descriptors by applying 4 statistical functionals, i.e., maximum, minimum, mean, and standard deviation, on a total of 26 X 9 trajectories - forming the final set of 936 features per frame.

The basic idea of our newly-proposed features is to essentially track spectral energy changes within a long-durational frame in both the directions of frequency-axis (spatial) and time-axis. Since the framework is inspired from video descriptor's extraction approach, the physical meaning related to speech production/perception can be difficult to establish. However, this framework provides a straightforward approach to quantify various inter-relationship between spectral-temporal characteristics in the speech signal directly from the time-frequency representations without much higher-level processing.

3. Experimental Setup and Results

In this work, we conduct the following two experiments on the emotion recognition tasks mentioned in section 2.1:

- **Exp I:** Comparison and analysis of our proposed *Traj-ST* with *Conv-PS* and *OpEmo-Utt* features in the three emotion recognition tasks
- **Exp II:** Analysis of recognition accuracy after fusion of *Traj-ST* with *Conv-PS* and/or *OpEmo-Utt* features in the three emotion recognition tasks

The *Conv-PS* feature extraction approach is similar to the *Traj-ST*, but instead of computing spatial-temporal characteristics on trajectories of Mel-filter bank output, we compute 45 low-level descriptors including fundamental frequency (f0), intensity (INT), MFCCs, their delta, and delta-delta every 10ms. We then derive the frame-level features by applying the 7 statistical functionals (max, min, mean, standard deviation, kurtosis, skewness, inter-quantile range) on these LLD features. This results in a total of 315 features per frame for Conv-PS. *OpEmo-Utt* is an exhaustive utterance-level feature set (i.e., emoLarge.config in the Opensmile toolbox) that has been used across many paralinguistic recognition tasks (Schuller *et al.*, 2013; Schuller *et al.*, 2014). It includes 6668 features in total per utterance. All features are znormalized with respect to an individual speaker. All evaluation is done via leave-one-speaker-out cross validation, and the accuracy is measured in unweighted average recall. Univariate feature selection based on ANOVA test is carried out for both *Traj-ST* and *Conv-PS* feature sets.

3.1 Recognition Framework

In Exp I, for *Traj-ST* and *Conv-PS* feature sets, we use Gaussian Mixture Model (M = 32) to generate a probabilistic score, $p_{i;t}$, for each class label at the frame-level, and then we perform utterance-level recognition using the following simple rule:

$$\underset{i \in \text{classes}}{\arg\max} \sum_{t=1}^{N} p_{i,t}$$

where *i* refers to the class label, *t* refers to the frame index, and N refers to the total number of frames in an utterance. For *OpEmo-Utt*, since it is a large-dimensional utterance-level feature vector, we utilize the GMM-based method after performing principal component analysis (90% of variance) and also linear-kernel support vector machine multi-class classifier.

In Exp II, the fusion methodology of *Traj-ST* with *Conv-PS* and *OpEmo-Utt* is depicted in Figure 2. The fusion framework is based on logistic regression. For *Traj-ST* and *Conv-PS*, the fusion is operated on the statistical functionals, i.e., mean, standard deviation, max, and min, applied on the $p_{i;t}$; and for *OpEmo-Utt*, the fusion is operated on the decision scores outputted from the one-vs-all multiclass support vector machine.

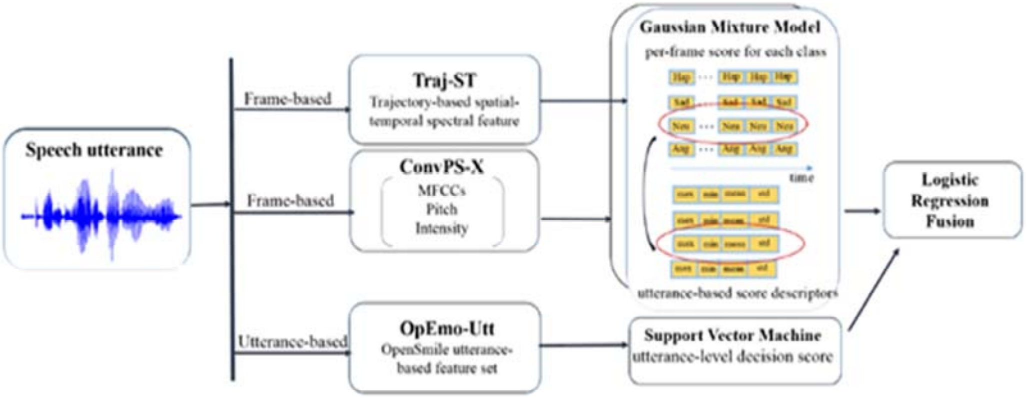

Speech utterance
Frame-based
Frame-based
Utterance-based
Traj-ST
Trajectory-based spatial-temporal spectral feature
ConvPS-X
MFCCs
Pitch
Intensity
OpEmo-Utt
OpenSmile utterance-based feature set
Gaussian Mixture Model
per-frame score for each class
Hap
Sad
Neu
Ang
time
utterance-based score descriptors
Support Vector Machine
utterance-level decision score
Logistic
Regression
Fusion

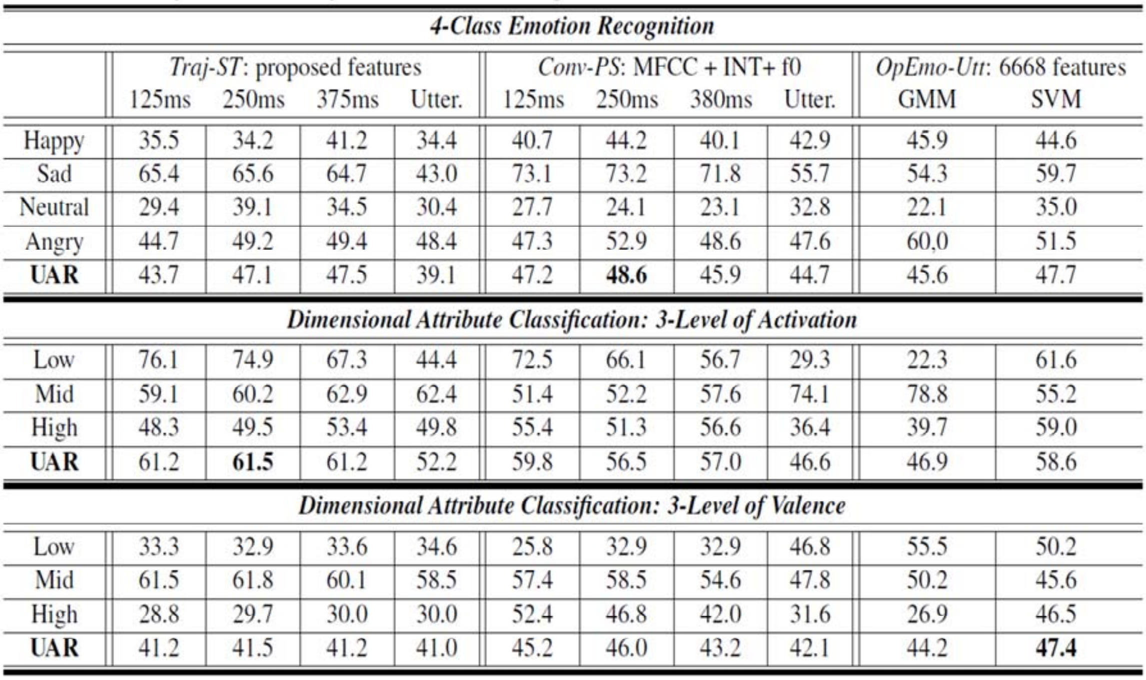
Figure 2. It depicts the fusion method for the three feature sets. Frame-based features are fused using statistical functionals of probabilistic scoring outputted from GMM model, utterance-level features are fused using decision score directly from the SVM classifier.The final fusion model utilized is logistic regression.

3.2 Exp I: Results and Discussions

Table 1 summarizes the detailed results of Exp I. For *Traj-ST* and *Conv-PS*, we report UARs of GMM model with different frame-length utilized for feature extraction, i.e., 125ms, 250ms, 375ms, and full-utterance length. For *OpEmo-Utt*, we report both UARs on using GMM and SVM models.

There are a couple points to note in the results. In the four-class emotion recognition task, *Traj-ST* compares comparably to *OpEmo-Utt* (47.5% vs. 47.7%), while the best accuracy is achieved by *Conv-PS* (48.6%). In the three-level valence recognition tasks, the best accuracy achieved is by using *OpEmo-Utt* (47.4%), where *Traj-ST* and *Conv-PS* do not perform well. Lastly, our proposed *Traj-ST* feature set performs significantly better than both *Conv-PS* and *OpEmo-Utt* on the task of three-level activation recognition. It achieves a recognition rate of 61.5%, which is an 1.7% improvement absolute over *Conv-PS* and 2.9% over *OpEmo-Utt*. By running the three types of emotion recognition tasks, it seems to be evident that each set of these features indeed possess a distinct amount and quality of emotional contents. *OpEmo-Utt* seems to perform the best for valence, possibly due to the complex nature on the perception of the degree of valence (i.e., requiring exhaust features to be extracted at the utterance-level). Although it has been demonstrated in the past that acoustic-related features often encodes more information in the activation dimension (Yildirim *et al*., 2004), it is quite still promising that see our proposed features, *Traj-ST*, are even more effective in predicting the overall perception of activation than these two other feature sets.

Table 1. It summarizes the detailed results of Exp I for three different emotion recognition tasks: 4-class emotion recognition, 3-level activation/valence recognition. For Traj-ST and Conv-PS, we report UARs of GMM model with different frame-length utilized for feature extraction. For OpEmo-Utt, we report both UARs on using GMM and SVM models.



	Traj-ST: proposed features				*Conv-PS*: MFCC + INT+ f0				*OpEmo-Utt*: 6668 features	
	125ms	250ms	375ms	Utter.	125ms	250ms	380ms	Utter.	GMM	SVM
4-Class Emotion Recognition										
Happy	35.5	34.2	41.2	34.4	40.7	44.2	40.1	42.9	45.9	44.6
Sad	65.4	65.6	64.7	43.0	73.1	73.2	71.8	55.7	54.3	59.7
Neutral	29.4	39.1	34.5	30.4	27.7	24.1	23.1	32.8	22.1	35.0
Angry	44.7	49.2	49.4	48.4	47.3	52.9	48.6	47.6	60,0	51.5
UAR	43.7	47.1	47.5	39.1	47.2	**48.6**	45.9	44.7	45.6	47.7
Dimensional Attribute Classification: 3-Level of Activation										
Low	76.1	74.9	67.3	44.4	72.5	66.1	56.7	29.3	22.3	61.6
Mid	59.1	60.2	62.9	62.4	51.4	52.2	57.6	74.1	78.8	55.2
High	48.3	49.5	53.4	49.8	55.4	51.3	56.6	36.4	39.7	59.0
UAR	61.2	**61.5**	61.2	52.2	59.8	56.5	57.0	46.6	46.9	58.6
Dimensional Attribute Classification: 3-Level of Valence										
Low	33.3	32.9	33.6	34.6	25.8	32.9	32.9	46.8	55.5	50.2
Mid	61.5	61.8	60.1	58.5	57.4	58.5	54.6	47.8	50.2	45.6
High	28.8	29.7	30.0	30.0	52.4	46.8	42.0	31.6	26.9	46.5
UAR	41.2	41.5	41.2	41.0	45.2	46.0	43.2	42.1	44.2	**47.4**

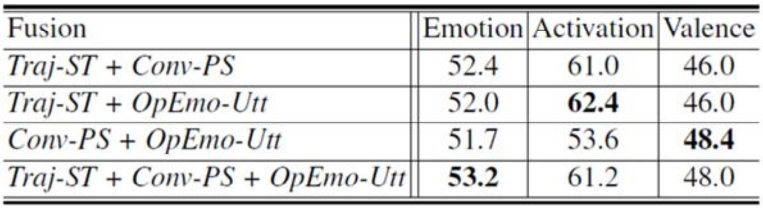
The frame duration also plays an important role in achieving the optimal accuracy for *Traj-ST* (also for *Conv-PS*). Our empirical finding seems to implicate that a duration of roughly 250ms is the optimal frame-duration - a result that corroborates findings in the previous use of long-term spectral features for emotion recognition (Chaspari *et al*., 2014; Chi *et al*., 2012). Furthermore, the feature selection output from *Traj-ST* shows that the top three directions of spatial-temporal characteristics are the {0,0} - base-trajectory, {1,0} - higher-spatial-equivalent-temporal directional trajectory, and {1,-1} - higher-spatial-earlier-temporal directional trajectory. These three constitutes 50% of the selected features. It is interesting to see that modeling not just the temporal changes but also the spatial (i.e., in the direction of frequency) can be beneficial; in specific, additional investigation will also need to be carried out to understand the reason to the finding that there seems to be a higher emotional discriminability in these specified trajectories, which quantify the spectral energy changes in direction toward higher-frequency bands.

In summary, we show that our novel feature set compares comparably to the state-of-art usage of exhaustive feature extractions in discrete 4-class categorical emotion recognition and outperforms significantly in the 3-level activation recognition.

3.3 Exp II: Results and Discussions

Given that in Exp I, each set of features seem to be capable of recognizing different representation of emotions. A natural experiment is to fuse the three different set of features. Table 2 lists the various fusion results. *OpEmo-Utt* refers to fusing the outputted decision scores from the SVM model.

Table 2. Exp II summary results on fusion of three different feature sets: Traj-ST, Conv-PS, OpEmo-Utt. The number presented is computed using UAR



Fusion	Emotion	Activation	Valence
Traj-ST + Conv-PS	52.4	61.0	46.0
Traj-ST + OpEmo-Utt	52.0	**62.4**	46.0
Conv-PS + OpEmo-Utt	51.7	53.6	**48.4**
Traj-ST + Conv-PS + OpEmo-Utt	**53.2**	61.2	48.0

There are a couple observations to be made with the results. The first is that fusion of different feature sets all improves the best single-feature set's result. In specifics, the best fusion accuracy of 4-class emotion recognition is 53.2% (4.6% absolute improvement over the best single-feature set) obtained by fusing all three sets of features; the best fusion result for 3-level valence is 48.4% (1.0% absolute improvement over the best single-feature set, *OpEmo-Utt*); lastly, the best fusion result for 3-level activation is 62.4% (0.9% absolute improvement over the best single-feature set, *Traj-ST*). We see that our newly propose features, *Traj-ST*, are indeed capable of additionally improve the recognition rate for categorical emotion recognition and activation level detection under this fusion framework - signifying the complementary information of our features possess with regard to emotional content that is originally lacking in these two well-established state-of-arts feature sets.

In summary, we have demonstrated that our novel trajectory-based spatial-temporal spectral features can be utilized in combination with the two popular and well-established acoustic feature sets in order to obtain improved emotion recognition rate.

4. Conclusions

In this work, we propose a novel set of low-level acoustic features derived directly from the spectrograms in order to characterize the long-term spatial-temporal information of the speech signal. We carry out emotion recognition experiments on both categorical emotion attributes and dimensional representations using the proposed features. Our experiments show that the newly-proposed feature set compares comparably to the well-established low-level acoustic descriptors and state-of-the-art exhaustive feature extraction approach on the categorical emotion recognition, and it outperforms on the task of activation level recognition. Furthermore, by fusing these trajectory-based spatial-temporal features, it improves the overall emotion recognition accuracy. Overall, it is quite promising to see these low-level features do possess emotional discriminatory power beyond the exhaustive set of established acoustic parameters.

There are several future directions. One of the immediate future direction is that we observe these features do not possess enough modeling power of the valence dimension; one of the possible causes may due to the fact that the grid-based differential operator may only capture the spatial-temporal interrelationship locally, and the statistical functionals may not be enough to quantify the suprasegmental information. We will immediately extend this grid-based differential operator to incorporate a wider range both in time and in space with different scales to capture the valence-related acoustic properties. Secondly, one of our main goals is to minimize the effort of raw signal processing required as we derive these features. We will replace the MFB portion of spectral representation to even lower-level (or employ sparse representation to ensure robustness) time-frequency representation while maintaining low computational complexity. Lastly, on the longer term, once these (raw) low-level features are developed, they can be suitable inputs to deep learning algorithms to learn various hierarchical representations of speech acoustic that are relevant for emotion perception. The ability to robustly recognize emotion will continue to be at the fore-front of developing human-centric applications

Acknowledgment

Thanks to Ministry of Science and Technology (103-2218-E-007-012-MY3) for funding.

References

Bach-y Rita, P. & Kercel, S. W. (2003). Sensory substitution and the human-machine interface. *Trends in cognitive sciences*, *7*(12), 541-546. doi:10.1016/j.tics.2003.10.013

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S.,...Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335-359. doi: 10.1007/s10579-008-9076-6

Calvo, R. A., D'Mello, S., Gratch, J. & Kappas, A. (2014). *The Oxford handbook of affective computing*. Oxford, England: Oxford University Press.

Campbell, W. M., Sturim, D. E. & Reynolds, D. A. (2006). Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308-311. doi: 10.1109/LSP.2006.870086

Chaspari, T., Dimitriadis, D. & Maragos, P. (2014). Emotion classification of speech using modulation features. In *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, 1552-1556.

Chi, T.-S., Yeh, L.-Y. & Hsu, C.-C. (2012). Robust emotion recognition by spectro-tempora modulation statistic features. *Journal of Ambient Intelligence and Humanized Computing*, *3*(1), 47-60. doi: 10.1007/s12652-011-0088-5

Childers, D. G. & Wu, K. (1991). Gender recognition from speech. Part ii: Fine analysis. *The Journal of the Acoustical society of America*, *90*(4), 1841-1856. doi: 10.1121/1.401664

Dobry, G., Hecht, R. M., Avigal, M. & Zigel, Y. (2011). Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 1975-1985. doi: 10.1109/TASL.2011.2104955

Eyben, F., Wöllmer, M. & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459-1462. doi: 10.1145/1873951.1874246

Hogan, N., Krebs, H. I., Sharon, A. & Charnnarong, J. (1995). *U.S. Patent No. 5,466,213A*. Cambridge, MA: Massachusetts Institute Of Technology.

Hollinger, G. A., Georgiev, Y., Manfredi, A., Maxwell, B. A., Pezzementi, Z. A. & Mitchell, B. (2006). Design of a social mobile robot using emotion-based decision mechanisms. In *Proceedings of f the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3093-3098. doi: 10.1109/IROS.2006.282327

Kim, Y., Lee, H. & Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3687-3691. doi: 10.1109/ICASSP.2013.6638346

Lee, C. M. & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293-303. doi: 10.1109/TSA.2004.838534

Lee, C.-C., Mower, E., Busso, C., Lee, S. & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, *53*(9), 1162-1171. doi: 10.1016/j.specom.2011.06.004

Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I.,...Sahli, H. (2013). Hybrid deep neural network-hidden markov model (dnn-hmm) based speech emotion recognition. In *Proceedings of 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 312-317. doi: 10.1109/ACII.2013.58

Mairesse, F. & Walker, M. (2006). Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 543-548.

Mower, E., Matarić, M. J. & Narayanan, S. (2011). A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(5), 1057-1070. doi: 10.1109/TASL.2010.2076804

Narayanan, S. & Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. In *Proceedings of IEEE Inst Electr Electron Eng.*, *101*(5), 1203-1233. doi: 10.1109/JPROC.2012.2236291

Nwe, T. L., Foo, S. W. & De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, *41*(4), 603-623. doi: 10.1016/S0167-6393(03)00099-2

Oneata, D., Verbeek, J. & Schmid, C. (2013). Action and event recognition with fisher vectors on a compact feature set. In *Proceedings of ICCV '13 Proceedings of the 2013 IEEE International Conference on Computer Vision*, 1817-1824. doi: 10.1109/ICCV.2013.228

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT press.

Scherer, K. R. (2003). Vocal communication of emotion: A review ofresearch paradigms. *Speech communication*, *40*(1-2), 227-256. doi: 10.1016/S0167-6393(02)00084-5

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J.,...Aharonson, V. (2007). The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *Proceedings of INTERSPEECH 2007*, 2253-2256.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C. & Narayanan, S. (2013). Paralinguistics in speech and languagestate-of-the-art and the challenge. *Computer Speech & Language*, *27*(1), 4-39. doi: 10.1016/j.csl.2012.02.005

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F.,...Zhang, Y. (2014). The interspeech 2014 computational paralinguistics challenge: cognitive & physical load. In *Proceedings of INTERSPEECH 2014*, 427-431.

Swartout, R. W., Gratch, J., Hill Jr, R. W., Hovy, E., Marsella, S., Rickel, J. & Traum, D. (2006). Toward virtual humans. *AI Magazine*, *27*(2), 96-108. doi: 10.1609/aimag.v27i2.1883

Vinciarelli, A., Pantic, M. & Bourlard, H. (2009). Social signal processing :Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743-1759. doi: 10.1016/j.imavis.2008.11.007

Wang, H., Kläser, A., Schmid, C. & Liu, C.-L. (2011). Action recognition by dense trajectories. In *Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3169-3176. doi: 10.1109/CVPR.2011.5995407

Wang, H., Kläser, A., Schmid, C. & Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, *103*(1), 60-79. doi: 10.1007/s11263-012-0594-8

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z.,...Narayanan, S. S. (2004). An acoustic study of emotions expressed in speech. In *Proceedings of INTERSPEECH 2004*, 2193-2196.