Evaluating speech content using remote sound sensing techniques

Panagiotis Zervas, Efthimios Bakarezos, Ara Movsesian, Nektarios Papadogiannis

Department of Music Technology and Acoustics Engineering

Technological Educational Institute of Crete, Greece.

Summary

This paper presents evaluation results on the task of speaker emotion recognition from audio cues captured with the laser beam deflection method (LBDM). For training emotional speech classification models we employed the well-established approach of Support Vector Machines (SVM). The remote sound sensing took place in a large auditorium where an optically reflecting surface, made of transparent double plexiglass, was excited by a loudspeaker reproducing sound content from an emotional database of acted speech. Results shows that the LBDM can yield good results for the task of speaker emotion classification and provided a promising solution for various applications associated with speech.

1. Introduction

Developing intelligent audio information systems able to extract meta-information from speech queues, is considered as one of the most challenging tasks. Recently, researchers have employed auxiliary sensor solutions (such as throat-microphones [1]) for improving accuracy in speech recognition and speech enhancement tasks.

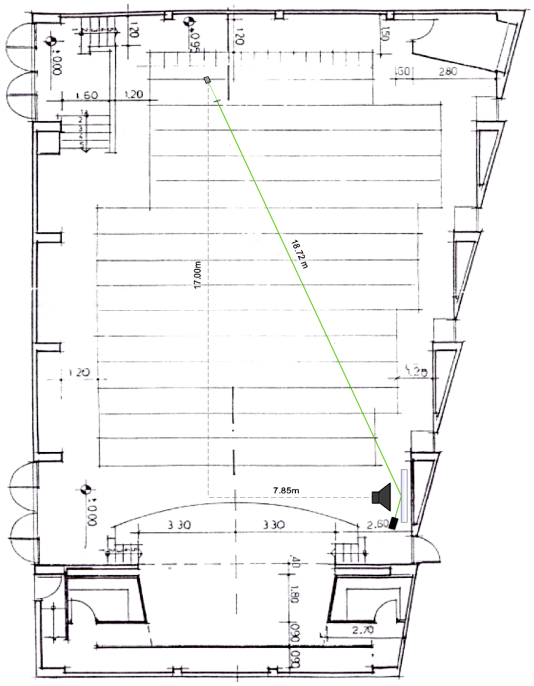
In this paper, we achieve remote capturing of speech using a sensor device based on the laser beam deflection method – LBDM. Generally, the method relies on the recording of laser beam path changes (deflection) resulting from its reflection of a vibrating object excited by a sound field [e.g. see 2] or from refractive index changes experienced when propagating in a medium in the presence of a sound field [e.g. see 3]. Our LBDM-based sensor is a non-contact, non-evasive measurement device capable of measuring vibration frequencies of distant objects. Specifically, this work initially focuses on the problem of remote recording of emotional speech using a simple LBDM-based device. Subsequently, the capability of capturing speech information with this device is evaluated for the task of speaker emotion classification. To this end we focus in the implementation of an emotion classification system able to perform robust, in situations where having a microphone is not an option (i.e. remote surveillance, emergency situations etc.).

Remote audio recording has been the subject of research investigation for many decades, due to the constraints set by the mechanical part of the microphones about their operation and sensitivity. Long distance recording of sound data using standard microphones is practically impossible, especially in situations where an object is between the speaker and the capturing device. Therefore, the research interest shifted to the development of optoelectronic devices, that with the utilization of light beams, can record and convert the vibrations caused to a material by acoustic waves, into audio data. These devices can achieve acceptable fidelity of recording over long distances due to the intense directionality of the beam and its intensity.

The paper is organized as follows. In Section 2, we describe the basic principles of our LBDM-based sensor in measuring acoustic speech signals and we detail our experimental setup. In Section 3, we formulate the problem of emotion recognition from speech. In Section 4, we describe the utilized emotional speech database and we give a brief description of the classification schema. Finally, in Section 5, we present experimental results that show the effectiveness of the proposed approach.

2. Experimental setup

All-optical remote recording of sound measurements took place in a large auditorium (Fig. 1).
An optically reflecting surface made of transparent double plexiglass (dimensions 0.8m × 1.1m, thickness 0.003m) was excited by a loudspeaker reproducing sound content from an emotional speech database. The beam from a continuous-wave (CW) laser (wavelength 532nm, maximum power 500mW) was incident on the reflecting surface at an



angle of 24.8°. The reflected laser beam was monitored using a commercially available CdS photodiode at a distance of 18.72m from the reflecting surface, using a lens of focal distance of 0.15m. The photodiode signal changed in way corresponding to the sound content exciting the reflecting surface, and the signal changes were recorded in a .wav format using appropriate software.

3. Emotion Classification Framework

Extensive research has been conducted lately on the task of emotion recognition from speech signals. Speech emotion classification is a key ingredient for building intelligent human-computer interaction interfaces for real-life applications relevant to robotics, gaming, health, well-being and other. An extensive number of experiments have been conducted using signal processing techniques to explore which aspects of speech would manifest saliently the emotional condition of a speaker. The outcome of this research was that the most crucial

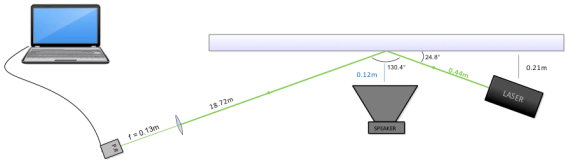


Figure 2: The implemented optoelectronic experimental device. PR - photoresist, f - convergent lens

aspects are those related to prosody [4], [5] (pitch contour, intensity and timing). Furthermore, voice quality [6] as well as certain co-articulatory phenomena [7] is also high correlated with some emotional states.
Our research focuses on recognizing basic emotional conditions where their differentiation is evident. In particular, our data capture the emotional states of anger, fear, joy, sad and a neutral.

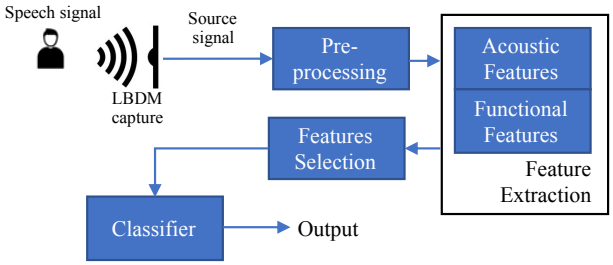


Figure 3: Classification framework diagram

Figure 3 depicts the whole process of the classification framework used in this study. First, an acted emotional speech database is used, which consists of annotated utterances. Next, the speech signal is captured from the LBDM module and pre-processed for removing reverberation artefacts as well as resampling and chunking the initial recordings. In a next step feature extraction is performed by using an open source feature extractor. Then, feature selection method is used and recognition is performed by a supervised classification algorithm.

3.1 Emotional Speech Data

Obviously, the main issue facing researcher initiating a study on emotional state recognition from speech, is the availability of appropriate data. Investigation of emotional states is possible with speech material obtained form: (a) spontaneous speech, (b) acted speech, or (c) elicited speech. All

Table 1: Feature set description

Low level acoustical features	
Cepstral (13)	MFCC 0 – 12
Spectral (35)	Mel spectrum bins 1-26 (0-8 kHz), zero crossing, 25%, 50%, 75%, and 90% spectral roll-off points, spectral flux, spectral centroid, relative position of spectral max & minimum
Energy	Log energy, energy in bands from 0-250Hz, 0-650Hz, 250-650Hz, 1-4 kHz, 3010 - 9123 Hz
Voicing	F0 (with cepstrum auto-correlation), F0 envelope, probability of voicing
Functional features	
Statistical	arithmetic mean, root quadratic means, root quadratic, geometric, arithmetic mean of non-zero values, number of non-zero values, zero crossing rate, centroid, variance, standard deviation, skewness, kurtosis, quartiles and inter-quartile ranges, 95 %, 99% percentile
Regression related	linear regression slope, offset, corresponding approx. error (quadratic and linear), quadratic regression coefficients a, b, and c, corresponding approximation error (quadratic and linear)
Minima/maxima related	range, position of max/min, difference min/max to arithmetic mean, number of peaks, mean distance between peaks, arithmetic mean of peaks, difference mean of peaks - mean

the three approaches have both advantages and disadvantages, and none of them could be considered as optimal for simulating real-world conditions. In our experimental setup, we used a database of Greek acted emotional speech [8]. The original dataset includes 51 utterances (single words, short sentences, long sentences and passages) of Greek speech recordings and their transcripts, encoded in a 2-channel interleaved 16-bit PCM.

3.2 Pre-processing

Upon completion of speech capturing with the LBDM approach the step of reverberation and noise removal was performed. In both cases the open source audio editing software of Audacity [9] was used. The configuration required for performing the reverberation step was studied separately for each emotional state due to the specificity of each emotion in the frequency, intensity, general in the perceptual characteristics of speech and its behaviour in a large resonance space. Finally, a common profile was used for all emotional states that best performs on each of them. A similar technique was followed in creating the noise profile as in the reverberation removal algorithm, in particular, an audio file was selected from which the profile of the existing noise was estimated.

3.3 Feature Extraction

Performance of any emotion recognition strategy largely depends on how relevant features, invariant to speaker, language, and contents could be extracted. Previous research in the field of emotion recognition has shown that emotional reactions are strongly related to the pitch and energy of the spoken message [10]. For example, the pitch of speech associated with anger or happiness is always higher than that associated with sadness or fear, and the energy associated with anger is greater than that associated with fear.

In this study, the feature extraction tool used was the OpenSmile (Open Speech and Music Interpretation by Large Space Extraction) [11].

OpenSmile provides configuration files that can be used for extracting predefined set of features. In particular we adopted the emo_large.conf configuration file which extracts a total of 6669 features. From this 57 are low level descriptors, 39 functional descriptors with the addition of their delta and delta-delta coefficients.

From the 57 low level descriptors, 39 functionals are computed after adding delta and delta-delta coefficients, resulting in a total number of 6669 features. A brief description of the extracted feature set is listed in Table 1 and include values such as mean, standard deviation, percentiles and quartiles, linear regression functionals, or local minima/maxima related functionals. In combination with the low-level descriptors, the comprehensive set of functionals embodies an extensive characterization of the audio signal.

3.3 Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model construction. The key assumption when using a feature selection technique is that the data contain many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context [12].

To deal with this issue, we used the Correlation based Feature Selection (CFS) [13] method from

the Weka toolkit [14]. CFS is based on the hypothesis that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other.

3.4 Classification

Once the features were extracted, standardized and selected, they were used to form the feature vector database. Each data sample in the database is an instance used for classification. Since each instance is labeled with the appropriate emotion, supervised classification algorithms were used. Many factors motivate the consideration of diverse classifications methods: tolerance to high dimensionality, capability of exploiting sparse data, and handling of skewed classes. Linear Discriminant Classifiers (LDCs) and k-Nearest Neighbour (kNN) classifiers are popular since the very first studies. They turned out to be efficient for acted and non-acted speech but show problems with the increasing number of features that leads to regions of the feature space where data is very sparse. Also, well known, Support Vector Machines (SVM) [15] is a natural extension of LDCs which provides good generalization properties even for a large feature vector. In our experiments the well-established approach of SVM was employed.

4. Experiments

4.1 Experimental setup

The experiments were conducted in the following order. First, the OpenSmile feature extraction tool extracted 6655 features for both emo_Clear and emo_LBDM recordings. Next, all of the extracted features were ranked using the CFS approach. The resulted datasets were used for training the SVM models. As an evaluation metric, we used the fold validation technique. A 5-fold validation approach was utilized and the utterances composing the train-test sets of each fold were hand picked.

4.2 Results

To evaluate the recognition results from the proposed machine learning framework, human evaluation was conducted on both the clean emotional database and the LBDM recordings.

In the context of human evaluation 10 listeners who had not previously heard the recordings where asked to categorize the recordings in the 5 emotional states. Since the database was composed of acted speech from a single female speaker, clear data evaluation gave a fairly high recognition rate of 98%. On the other hand, evaluation of LBDM recordings showed a reduction in the classification with a 78% total accuracy. It is worth noting that a noticeable number of listeners were able to classify the emotional state correctly even in cases of recordings with very low intelligibility, which

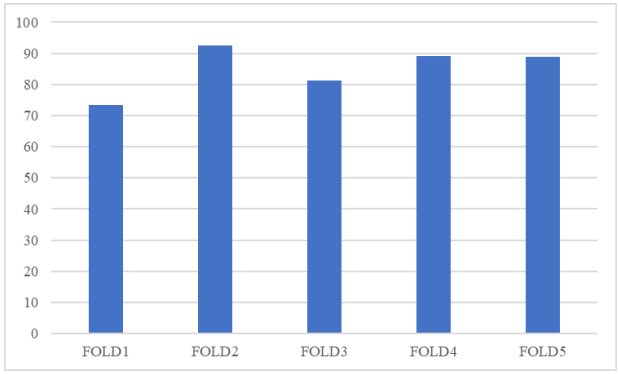


Figure 4: Total accuracy of LBDM signal emotion classification

proves the preservation of prosodic features in the LBDM recording. Even in conditions of intense loudness and low intelligibility, emotional state was particularly clear especially in the categories of anger, joy and neutral.

To evaluate the ability to identify emotions from

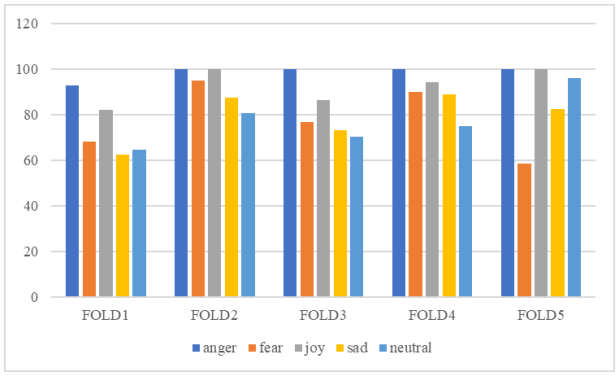


Figure 5: Per class accuracy on LBDM signals

LBDM captured speech, 5-fold cross-validation was employed. Table 2 shows the total accuracy for each of the experiments carried out as well as the human evaluation results.

Figure 4 depicts the total accuracy achieved with LBDM signal training for each fold. Finally, figure 5 illustrates the classification results for each emotional class. It shows that anger and joy had the highest accuracy while fear was the most difficult emotional state to identify.

Table 2: Emotion classification total accuracy

Dataset	Evaluation	Total Accuracy
Clear Data	Human	98
	SVM	96
LBDM Data	Human	78
	SVM	85

At this point, it should be noted that human evaluation of LBDM data showed lower total accuracy compared to the SVM classification model. This may in part be justified by the large number of features that have been used to create the training vector but may be an indication of models overfitting due to the small number of training data.

5. Conclusions

In this paper, we achieve remote capturing of speech using a sensor device based on the laser beam deflection method capable of measuring vibration frequencies of distant objects. For the purpose for evaluating the proposed a approach as an audio sensing device we conducted experiments for the classification of speaker's emotional state. Results showed that LBDM can be applied with sufficient success for the problem of identifing speaker emotional state. Therefore, this novel sensor and signal processing method is expected to provide a promising solution for various applications associated with speech. In our future research plans is the study of procedure performance in classification scenarios with more speakers, real-life scenarios and with different approaches in capturing and pre-processing the recorded LBDM signal.

References

[1] E. Erzin, "Improving Throat Microphone Speech Recognition by Joint Analysis of Throat and Acoustic Microphone Recordings," in IEEE Trans. on Audio, Speech, and Language Proc., vol.17, no.7, pp. 1316-1324, Sept. 2009.

[2] K. Roark, G.J. Diebold: Resonant microphone based on laser beam deflection. Journal of Applied Physics 96 (2004) 864-866.

[3] S.M. Maswadia, B.L. Ibey, C.C. Roth, D.A. Tsyboulski, H.T. Beier, R.D. Glickmang, A.A. Oraevsky: All-optical optoacoustic microscopy based on probe beam deflection technique. Photoacoustics 4 (2016) 91-101

[4] Murray, I. R., and Arnott, J. L. (1993). "Toward the simulation of emotion in synthetic speech—A review of the literature on human vocal emotion," J. Acoust. Soc. Am. 93, 1097–1108.

[5] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, Philadelphia, PA, 1996, pp. 1989-1992 vol.3.

[6] Kienast, Miriam / Sendlmeier, Walter F. (2000): "Acoustical analysis of spectral and temporal changes in emotional speech", In SpeechEmotion-2000, 92-97.

[7] H. Fujisaki, K. Hirose, Analysis of voice fundamental frequency contours for declarative sentences of Japanese J. Acoust. Soc. Jpn. (E) 5, 4 (1984)

[8] Fakotakis, N.: Corpus Design, Recording and Phonetic Analysis of Greek Emotional Database. In: Proceedings of the International Conference on language Resources and Evaluation (LREC), pp. 1391–1394 (2004)

[9] Audacity [Open Source audio editing software] (2017).

[10] Tatham M., Morton K., Expressions in Speech: Analysis & Synthesis, Oxford Lingiistics, (2004)

[11] Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller: "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013.

[12] Mohd Syahid Anuar, Ali Selamat, and Roselina Sallehuddin, Particle swarm optimization feature selection for violent crime classification, pp. 97–105, Springer International Publishing, Cham, 2014.

[13] M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

[14] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[15] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning. 20 (3): 273–297.