EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

EDITORIAL Open Access

Emotion and mental state recognition from speech

Julien Epps[1,2*], Roddy Cowie[3], Shrikanth Narayanan[4], Björn Schuller[5] and Jianhua Tao[6]

As research in speech processing has matured, attention has gradually shifted from linguistic-related applications such as speech recognition towards paralinguistic speech processing problems, in particular the recognition of speaker identity, language, emotion, gender, and age. Determination of a speaker's emotion or mental state is a particularly challenging problem, in view of the significant variability in its expression posed by linguistic, contextual, and speaker-specific characteristics within speech. In response, a range of signal processing and pattern recognition methods have been developed in recent years.

Recognition of emotion and mental state from speech is a fundamentally multidisciplinary field, comprising contributions from psychology, speech science, linguistics, (co-occurring) nonverbal communication, machine learning, artificial intelligence and signal processing, among others. Some of the key research problems addressed to date include isolating sources of emotion-specific information in the speech signal, extracting suitable features, forming reduced-dimension feature sets, developing machine learning methods applicable to the task, reducing feature variability due to speaker and linguistic content, comparing and evaluating diverse methods, robustness, and constructing suitable databases. Studies examining the relationships between the psychological basis of emotion, the effect of emotion on speech production, and the measurable differences in the speech signal due to emotion have helped to shed light on these problems; however, substantial research is still required.

Taking a broader view of emotion as a mental state, signal processing researchers have also explored the possibilities of automatically detecting other types of mental state which share some characteristics with emotion, for example stress, depression, cognitive load, and 'cognitive epistemic' states such as interest, scepticism, etc. The recent interest in emotion recognition research has seen applications in call centre analytics, human-machine and human-robot interfaces, multimedia retrieval, surveillance tasks, behavioural health informatics, and improved speech recognition.

This special issue comprises nine articles covering a range of topics in signal processing methods for vocal source and acoustic feature extraction, robustness issues, novel applications of pattern recognition techniques, methods for detecting mental states and recognition of non-prototypical spontaneous and naturalistic emotion in speech. These articles were accepted following peer review, and each submission was handled by an editor who was independent from all authors listed in that manuscript. Herein, we briefly introduce the articles comprising this special issue.

Trevino, Quatieri and Malyska bring a new level of sophistication to an old problem, detecting signs of depressive disorders in speech. Their measures of depression come from standard psychiatric instruments, Quick Inventory of Depressive Symptomatology and Hamilton Depression rating scales. These are linked to measures of speech timing that are much richer than the traditional global measures of speech rate. Results indicate that different speech sounds and sound types behave differently in depression, and may relate to different aspects of depression.

Caponetti, Buscicchio and Castellano propose the use of a more detailed auditory model than that embodied in the widely employed mel frequency cepstral coefficients, for extracting detailed spectral features during emotion recognition. Working from the Lyon cochlear model, the authors demonstrate improvements on a five-class problem from the speech under simulated and actual stress database. Their study also further validates the applicability of long short-term memory recurrent neural networks for classification in emotion and mental state recognition problems.

Callejas, Griol and López-Cózar propose a mental state prediction approach that considers both speaker

* Correspondence: j.epps@unsw.edu.au
[1]School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia
Full list of author information is available at the end of the article

intentions and emotions in a spoken dialogue context. Their system comprises an emotion recognizer that relies on the user's speech and dialogue history, and an intention recognizer that relies on the user input and the system's prompt. Experiments were carried out using an academic information dialogue system, and validated with both simulated and real users, and results show improved interaction efficacy.

Tao et al. present a method for providing an audio-visual mapping between neutral speech content and neutral facial expression that is then used to contrast with observed, emotional audio-video data for emotion recognition. This approach yields a degree of independence between the facial expression and the uttered expression, which are known to exhibit conflicting information during some instances of spontaneous face-to-face communication. The method provides advantages over four more conventional bimodal emotion recognition methods, when evaluated on a six-emotion semi-natural audio-video database.

An investigation of the effect of cognitive load on a variety of vocal tract parameters is reported by Yap, Epps and Ambikairajah. Their analysis of two databases, comprising a laboratory-style Stroop task and a more naturalistic reading comprehension task, finds that formant frequencies are affected by cognitive load, in particular the slope, range and duration of the formant trajectory. In classification experiments, the first three formant frequencies were found to provide comparable or improved performance relative to the Mel frequency cepstral coefficients, which have a much higher feature dimension.

Pon-Barry and Shieber propose the use of phrase-level prosodic features for inferring the level of a speaker's self-reported uncertainty. Following a description of methods for creating an 'uncertainty in speech' corpus, the authors study the differences between self-reported uncertainty and externally perceived uncertainty. Their investigation of the effects of uncertainty on a range of prosodic features forms the basis for an uncertainty classification system, which achieves encouraging accuracies that are further increased when phrase-level prosodic features are used in combination with prosodic features from 'target words' elicited using their methods.

The article by Weninger et al. describes the application of non-negative matrix factorization, employed with success in robust speech recognition, to noisy emotion recognition. Experiments on the FAU Aibo Emotion Corpus degraded by babble and street noise show that the technique is promising, particularly for realistic conditions of low signal-to-noise ratio and for mismatched training and test data.

Hansen, Kim and Rahurkar demonstrate examples of substantial improvements to emotion recognition by subband weighting of the Teager energy operator-based critical band autocorrelation envelope feature. The empirically derived subband weights were also found to reduce speaker dependency, when evaluated on a novel speech under stress database with physiological ground truth.

Finally, Iliev and Scordilis report details of a method for estimating the symmetry of the glottal pulse, and the extent to which it can characterise emotion in speech. Their method, which extracts information on the relative durations of the glottal opening and closing phases from inverse filtered speech, performs very effective classification for a single-dimension feature, when evaluated on a six-emotion-acted speech database.

Through these articles, this issue provides avenues for recognizing mental states more robustly in disturbed audio feeds and with less dependence on the speaker and facial expression. This is obtained by auditory modelling, improved or novel feature types, phone-specific information, spectral decomposition and context exploitation. The articles further cover a broad range of affective states, reaching from depression disorders to cognitive load, and general emotion.

Acknowledgements

The guest editors would like to thank the EURASIP JASP Editor-in-Chief, Prof. Phillip Regalia, for the opportunity to coordinate this special issue, and the anonymous reviewers for their diligent work, without which this special issue would not have been possible.

Author details

[1]School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW 2052, Australia [2]ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh, NSW 2015, Australia [3]Queen's University, Belfast BT7 1NN, Northern Ireland [4]Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA [5]Institute for Human-Machine Communication, Technische Universität München, 80290 München, Germany [6]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China

Competing interests

The authors declare that they have no competing interests.

Received: 8 December 2011 Accepted: 19 January 2012
Published: 19 January 2012

doi:10.1186/1687-6180-2012-15
Cite this article as: Epps *et al.*: Emotion and mental state recognition from speech. *EURASIP Journal on Advances in Signal Processing* 2012 2012:15.