Measuring Affective and Social Signals in Vocal Interaction

Khiet P. Truong
Human Media Interaction, University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
k.p.truong@ewi.utwente.nl

ABSTRACT
In this paper, I will discuss how and what type of measurements of vocal interactional behavior can be used to recognize affective and social signals. Three studies will be presented that deal with 1) the collection and recognition of spontaneous vocal and facial expressions in a gaming context, 2) the detection of laughter in meetings, and 3) the relation between dominance and overlapping speech in multiparty conversations. On the basis of these studies, (dis-)advantages and issues in speech processing for affective and social computing will be evaluated. Acoustic features, but also simple speech or no-speech information were employed in these studies. In addition, fundamental issues such as 'ground truth labeling' and collection of spontaneous data are also discussed.

Author Keywords
speech, social signals, affective computing.

ACM Classification Keywords
H5.1. Multimedia information systems: Audio input/output
H5.2. User Interfaces: Natural language

INTRODUCTION
In speech analysis research, there is a long history of analysing speech signals with respect to human affective and social signals. One of the first studies carried out on real-life, natural emotion speech material [1], described an acoustic analysis of a well-known radio broadcast of the explosion of the Hindenburg Zeppelin in 1937. The audio sample contains the speech of a radio reporter who witnessed the explosion live during broadcasting and who continues his report in an emotional tone of voice. Speech analyses were performed by hand on a relative small amount of speech data. In the 70s, automatic speech feature extraction methods and statistical learning algorithms made it possible to develop simple automatic speech recognition systems (ASR). The tasks of these systems was to recognize *what* is said. Automatically recognizing *how* something is said gained more interest in the past 10-15 years, when the number of serious efforts into the automatic recognition and understanding of human affective and social signals was (and still is) growing steadily. I will briefly present three studies that we have carried out and discuss how the speech modality was used for our research. Not only speech-specific issues will be discussed but also some general issues in affective computing such as emotion labeling and the collection of spontaneous data.

SPEECH-BASED RECOGNITION OF AROUSAL AND VALENCE
An increasing number of researchers adopt the arousal-valence dimensional model of affect for automatic affect recognition. However, only a few databases containing spontaneous vocal expressions have been developed with continuous arousal-valence annotations. In addition, we are also interested in investigating differences between 'felt' affect annotations performed by people who have undergone the emotion themselves, and 'perceived' annotations performed by naive observers. Hence, we decided to record our own audiovisual spontaneous corpus of affect. We describe how we developed speech-based affect recognizers that are trained to predict a location in the arousal-valence space.

Collecting and Labeling Data: the TNO-Gaming Corpus
An audiovisual corpus containing expressive vocal and facial behavior was collected by inviting people to play a multiplayer videogame. The gaming sessions took place at TNO in Soesterberg, the Netherlands. In order to obtain affect annotations, each participant labeled his/her own affect after each gaming session. Seventeen males and eleven females with an average age of 22.1 years (2.8 standard deviation) participated in the gaming experiment. Vocal behavior was especially stimulated and was recorded by microphones that were attached near the mouths of the participants to reduce the effects of crosstalk (facial expressions were also recorded by webcams). The video content of the game itself was also stored. A more detailed description of this corpus can be found in [2].

Affect Rating by the Gamers Themselves ('Felt') and Naive Observers ('Perceived')

After each gaming session, the participants annotated their own emotions based on the video recordings and the videostream of the game content itself. We asked the participants to recall what they were *feeling* during playing. The participants rated the running video and could not pause or rewind the video. Two scales were used for rating, namely the arousal (active-passive) and valence (positive-negative) scale. Each 10 seconds, an arrow appeared on the screen to signal the participants to give an arousal and valence rating *separately* on a scale from 0 to 100. We will call these ratings the *self*-ratings. In this way, a total of 7473 affect-rated speech segments were obtained (after speech segmentation). Out of this total, 2400 segments were selected (sampling the whole arousal-valence space of the self-annotations evenly) for re-rating performed by 6 naive observers. Similar to the rating procedure of the gamers, the naive observers were asked to rate what they *perceived* on the arousal and valence scale. Each observer rated different parts of the dataset that overlapped with parts that were rated by other observers, such that each segment was rated by 3 different observers. In order to obtain unique ratings for each segment, the 3 different ratings were averaged. We refer to these ratings as *other.avg*-ratings.

Agreement Between Gamers and Naive Observers

To what extent do the ratings given by the gamers themselves differ from the ones given by the observers? To answer this question, the continuous ratings were discretized into 5 classes and Krippendorff's α (ordinal, [3]) was applied to assess the level of agreement between the *self*-ratings and the *other.avg*-ratings. Relatively low agreement scores were obtained: 0.27 and 0.36 for arousal and valence respectively (the agreement *among* the external observers was 0.28 and 0.57 for arousal and valence respectively). How does this discrepancy influence the performance of affect recognizers developed with these two types of ratings?

Recognition Task

Two types of speech-based arousal and valence recognizers were developed in parallel: one based on the self-ratings and the other based on the other.avg-ratings. The task of the affect recognizers was to predict scalar values on continuous scales of arousal and valence, rather than to recognize categories of emotions.

Method and Features

A Support Vector Regression approach was used to predict arousal and valence values (see [2]). As speech features, we extracted a selection of features that were commonly used in emotional speech research as described in the literature. First, a voiced-unvoiced detection algorithm (available in Praat [4]) was applied to find the voiced units. The features were extracted over each voiced unit of a segment. In addition, global information calculated over the whole segment (instead of per voiced unit) about the speech rate and the intensity and pitch contour was included. The following features were extracted: 4 pitch-related (mean, standard deviation, range (max-min), mean absolute pitch), 4 intensity-related (Root-Mean-Square, mean, range (max-min), standard deviation), 5 energy-distribution-in-spectrum-related (slope Long-Term Averaged Spectrum, Hammarberg index, standard deviation, centre of gravity, skewness), 24 Mel Frequency Cepstrum Coefficents (12 coefficients, 12 first order derivatives), and 6 other features (speech rate1, speech rate2, mean positive/negative slope pitch, mean positive/negative slope intensity). Subsequently, the features extracted on voiced-unit-level were aggregated to segment-level by taking the mean, minimum, and maximum of the features over the voiced units. Hence, we obtained per segment a feature vector with 117 dimensions. These features were normalized for speaker variation by transforming the features to z-scores: $z=(x-\mu)/\sigma$, with μ and σ calculated over a development set.

Results

Similarly to the human-human agreement assessment, the results are expressed in terms of Krippendorff's α (to allow for comparison): predicted values and reference ratings were first discretized into 5 classes, and Krippendorff's α was computed. We can observe in Table 1 that the recognizer based on *other.avg*-ratings performs much better than the *self*-based recognizer; the *other.avg*-ratings appear to describe the affect perceived more consistently.

	self	*other.avg*
arousal	0.22	0.42
valence	0.10	0.28

Table 1. Results recognition task, expressed in Krippendorff's α [3] (i.e., agreement between man-machine).

Discussion

What we consider 'ground truth' is of great influence on the performance of the recognizer. In this study, the *self*-ratings do not appear to be consistent enough for the learning algorithm, whereas the perceived affect ratings give better performance. It remains a challenge to acquire reliable affect ratings as the perception of affect is influenced by many factors such as context, culture and personality. With respect to automated recognition of affect, the acoustics-based recognizers have difficulty recognizing valence; a fusion with other modalities, such as facial expressions, could boost performance considerably. Finally, there appears to be an increasing interest for the recognition of gradations of affect. Adopting the dimensional approach and applying regression techniques to recognize these gradations is a relatively new approach, and hence, more investigation is needed into e.g., adequate evaluation procedures for these models.

AUTOMATIC DETECTION OF LAUGHTER IN MEETINGS

In this study, our goal was to detect paralinguistic events in speech (see also [5]). One the most recognizable events for humans, and (luckily) often annotated event is laughter. The first step towards this goal was to discriminate between laughter and speech. One could ask the question why developing a separate laughter detector is necessary as most ASR systems already have laughter models included. ASR systems are typically not tuned to detect paralinguistic events, in fact, these events are usually seen as 'garbage'. In addition, the computation cost and labour needed to train such a system would not make ASR a good candidate for the detection of paralinguistic events.

Data - ICSI Meeting Corpus

As speech data, the ICSI Meeting Corpus was used [6]. For training and testing, respectively 30 and 3 meetings were used, totaling an amount of approximately 90 minutes (>2500 segments) for each class of laughter and speech separately.

Method and Features

We extracted 12 Perceptual Linear Coding Coefficients (PLP, which model spectral properties of the speech according to a model adapted to the properties of the human ear) plus 1 energy component and their 1st order derivatives. After normalization to z-scores, these features were used in Gaussian Mixture Models (GMMs) to train a laughter and speech model. In classification, a log-likelihood ratio is used to decide the class.

Results

With relatively straightforward methods and features, an Equal Error Rate (EER, the point where the false alarm rate is equal to the miss rate) of approximately 6% was achieved. Note that only audible laughter segments were included in the speech material; e.g., unvoiced laughter was not part of this evaluation task.

Discussion

The next step would be to perform (real-time) laughter detection (e.g., [7]) and to use additional modalities (e.g., [8]). Furthermore, the detector does not give an interpretation of the laughter yet, and does not make distinctions between types of laughter (e.g., voiced vs unvoiced). Rather, we view the output of this detector as a useful feature for a higher-level affect recognition system.

ANALYSIS OF OVERLAPPING SPEECH AND DOMINANCE IN MULTIPARTY CONVERSATION

In social sciences and psychology, interuptions have frequently been studied with respect to cultural, gender, and status aspects. Traditionally, interruptions are treated as indicators of power, control, and dominance [12]. However, while some interruptions may indeed be seen as power displays, some of these are actually rapport displays. But in general, given the assumptions that turn-taking is regulated by the notion of 'one speaker speaks at a time', the social convention that it is impolite to speak at the same time when someone else is speaking, and the fact that it is difficult to decode the message when two speakers are speaking at the same time, interruptions are generally perceived as rude and competitive. In this study, we explored how interruptions can tell us something about social group dynamics in multiparty conversation, dominance and speaker role in particular.

Data - AMI Meeting Corpus

As multiparty conversation data, the AMI Corpus was used, see [9] for a detailed description. In short, the AMI corpus is comprised of recorded meetings in which 4 participants are brainstorming about the development of a tv remote control. Each participant has a role: there is a project manager, a user interface expert, a marketing expert, and an industrial designer. The corpus has multimodal and multi-layered annotations of e.g., gaze, head and hand gestures, and dominance. For our analyses, we used the manual word transcriptions and the dominance annotations as described in [10]. Each meeting (we used IS1000a, IS1001b, IS1003b, and IS1006b) was divided into 5-minute segments in which each meeting participant was ranked by dominance by 3 different raters. Only those segments were used where there was a majority agreement on ranked dominance.

Overlapping Speech Analysis

Overlapping speech parts were automatically found based on the manual speech transcriptions and the following definition of overlapping speech: we speak of overlap when there is more than 1 person talking at the same time. The person who performs the overlapping is the active overlapping speaker, while the overlappee is the passive overlapped speaker. We adopted the following measurements of vocal behavior as proposed by [10] who have analyzed overlapping speech in the context of interviews: an attack/resist ratio R and the overall frequency of active overlaps D. R is calculated as (A-P)/(A+P) where A is the number of words spoken by the active overlapper and P is the number of words spoken by the passively overlapped. A negative R indicates that the speaker is more a floor keeper than an overlapper. D is calculated as 100*A/(M+P+A) where M is the number of words spoken in a mono-speaker condition. A low D indicates low density of active overlaps. In [10], these measures were succesfully applied to journalists and interviewees: journalists had higher R and D than interviewees.

Results

The R and D measurements were compared to the dominance annotations which gave mixed results. In 2 of the 4 meetings analyzed, the measures appeared to show relation to dominance, see Figure 1. However, counter-intuitively, low R and D appeared to be associated with high dominance. In addition, correlating R and D to role (i.e., project manager etc.) did not reveal any patterns.

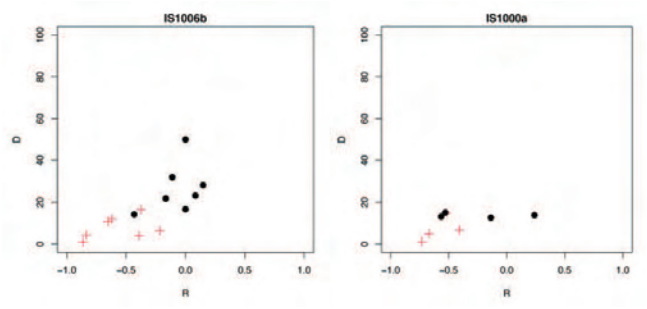

Figure 1. '+' is the most dominant person, '.' the least dominant, R on the x-axis, D on the y-axis.

Discussion

We have to keep in mind that the R and D measures are designed for a journalist-interviewee setting. However, the thought behind these measurements may be the same: namely, an overlapper is usually perceived as dominant. The fact that we did not obtain the expected values for R and D may have to do with the observation that in multiparty conversational speech there are a lot more cooperative interruptions, e.g., backchannels ('mm-hmm'), than in journalists-interviewees settings. Calculating these measures for competitive overlaps *only* may show improvement.

CONCLUSIONS AND DISCUSSION

We have given a brief overview of some analyses of affective and social signals in speech to show how different measures of vocal behavior can inform us about human affective and social behavior. Specifically, we looked at arousal/valence prediction, laughter detection, and a classic example of a social signal namely dominance. First, the study in arousal/valence prediction showed how differences in annotation affect the recognizer's performance, and the study showed that the collection of spontaneous affect data remains a challenge: for speech analysis, close-talk microphones that pick up few surrounding noise or talk are preferred (in the first study, the microphones were attached with tape to the face near the mouth), although in real-world situations, this is not always possible. We also looked at a more low-level feature of affect, namely laughter, for which relatively reliable annotations are available. In this case, relatively high detection performances were obtained with the disadvantage that the detector trained only detects laughter without giving an interpretation. Modeling the interpretation is not only a matter of collecting a sufficient amount of data for the training of algorithms, it is also about modeling the *context*. For example, depending on the situation, laughter can signal humor, joy, politeness or shyness. The first two studies covered the modeling of affect with acoustic features whereas in the third study, we looked at simple speech and no-speech information to model social signals such as dominance. With the on-and-off pattern of speech, overlaps in multiparty speech were analyzed to see whether certain overlap behavior could be related to social group behavior. In this case, information about the competitive or cooperative nature of the overlap in speech would help to reveal this relation. In general, speech as a modality offers an inobtrusive way of measuring affective and social signals. The use of speech with other measurements such as physiological measures is an interesting combination. Humans have high control over the vocal apparatus so that emotions in the voice can be suppressed, while it is much harder to suppress certain physiological measures. We also have to keep in mind that social conventions may pose limitations to what is expressed. For future research, we suggest to investigate how context and personality can be modelled for the interpretation of affect and social signals.

ACKNOWLEDGEMENTS

This research has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet).

REFERENCES

1. Williams, C.E. and Stevens, K.N. Emotions and speech: some acoustical correlates. *Journal of the Acoustical Society of America*, *52* (1972), 1238–1250.
2. Truong, K.P., van Leeuwen, D.A., Neerincx, M.A., and de Jong, F.M. Arousal and Valence Prediction in Spontaneous Emotional Speech: Felt Versus Perceived Emotion. In *Proceedings of Interspeech* (2009), 2027-2030.
3. Krippendorff, K. Computing Krippendorff's Alpha-Reliability (2007). Retrieved from http://www.asc.upenn .edu/usr/krippendorff/webreliability.doc.
4. Boersma, P. Praat, a system for doing phonetics by computer. *Glot International*, *5*, 9-10 (2001), 341-345.
5. Truong, K.P., and van Leeuwen, D.A. Automatic discrimination between laughter and speech. *Speech Communication*, *49* (2007), 144-158.
6. Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Macias-Guarasa, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., and Wrede, B. The ICSI meeting project: resources and research. In: NIST ICASSP2004 Meeting Recognition Workshop, Montreal, Canada, 2004.
7. Melder, W.A., Truong, K.P., den Uyl, M., van Leeuwen, D.A., Neerincx, M.A., Loos, L.R., and Plum, B.S. Affective multimodal mirror: sensing and eliciting laughter. In Proceedings of the International workshop on Human-Centered Multimedia (HCM) (2007), 31-40.
8. Reuderink, B., Poel, M., Truong, K. P., Poppe, R., and Pantic, M. Decision-Level Fusion for Audio-Visual Laughter Detection. In *Proceedings of MLMI*, (2008), 137-148.
9. Carletta, J. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus.

Language Resources and Evaluation Journal, 41, (2007), 181–190.

10. Hung, H., and Gatica-Perez, D. Identifying Dominant People in Meetings from Audio-Visual Sensors. In *Proceedings of FG 2008*, 1-6.
11. Adda-Decker, M., Barras, C., Adda, G., Paroubek, P., de Mareuil, P. B., Habert, B., Annotation and analysis of overlapping speech in political interviews. In *Proceedings of LREC* (2008), 3105-3111
12. West, C., Against our will: Male interruptions of females in cross-sex conversations. *Annals of the New York Academy of Sciences, 327* (1979), 81-97.