Verbal aggression detection in complex social environments

P.W.J. van Hengel

Sound Intelligence
Groningen, the Netherlands

T.C. Andringa

Department of Artificial Intelligence
University of Groningen
Groningen, the Netherlands

Abstract

The paper presents a knowledge-based system designed to detect evidence of aggression by means of audio analysis. The detection is based on the way sounds are analyzed and how they attract attention in the human auditory system. The performance achieved is comparable to human performance in complex social environments. The SIgard system has been deployed in a number of different real-life situations and was tested extensively in the inner city of Groningen. Experienced police observers have annotated ~1400 recordings with various degrees of shouting, which were used for optimization. All essential events and a small number of nonessential aggressive events were detected. The system produces only a few false alarms (non-shouts) per microphone per year and misses no incidents. This makes it the first successful detection system for a non-trivial target in an unconstrained environment.

1. Introduction

To improve the overall usefulness of camera-based surveillance systems it is important that situations with a high risk of injury and a relatively fast development, such as street-fights, are detected as quickly and as reliably as possible. Only then can appropriate proactive action be initiated. The ideal is an intelligent camera system that does not require a human observer to monitor the scene. This system must prioritize potentially dangerous situations autonomously and present high priority events to a human observer for a final check and possible follow-up.

Developments in image processing techniques are unlikely to realize this ideal, because visual cues are generally either insufficient or ambiguous. Furthermore, to ensure that all potentially dangerous situations will be in view of a camera, the number of cameras needs to be increased to an unrealistic level.

Of course, humans on the street do not only see but listen as well. In fact we use acoustical cues typically as an indication to raise alertness and to focus visual attention in the correct direction; audition guides vision.

This paper describes the development of the SIgard system that detects the presence of aggressive shouting in realistic, uncontrollable environments. This system must function autonomously in complex social environments such as a city center on a busy Saturday night, and present high priority events to a human observer for possible follow-up. Sound Intelligence has installed the verbal aggression detection system successfully in several Dutch city-centers, but also in prisons, in public transport (trains and train stations), and at various other sites.

This paper starts with a short overview of the problems associated with the development of the verbal aggression detection system and why a knowledge-based approach has been chosen over the standard statistical methods. The scientific basis of the model for verbal aggression is addressed in combination with the formulation of a number of cues. This leads to the design of an auditory scene analysis system with an attentional mechanism and tools to search for aggression cues in the attented signal. This system has been tested in a pilot-project and during the first 10 weeks of the official deployment in Groningen. The paper ends with a number of conclusions about the system itself and the reasons for its success in situations that are well outside the domain of traditional classification methods.

2. Unconstrained environments

There are two main problems associated with acoustic aggression detection. The first is the rarity of target events and the second the fact that the targets typically occur in environments in which a majority of much more frequent sounds, like laughter and or playful shouts, occur that are similar to verbal aggression. It is essential that these sounds do not lead to a false-alarm rate that is unacceptable for the user.

There are numerous well-known standard classification methods for acoustic signals, such as Neural Network, Bayesian Nets, and Hidden Markov Models. These *statistical* methods rely on the availability of training databases that are representative for the test and eventual deployment condition: the more data and the narrower the domain the better the classification results will be.

These methods are unsuitable for verbal aggression detection in unconstrained environments [1]. A first reason is that it is very difficult to acquire realistic training examples of the target event (acquiring a few realistic samples may require many months of recording and manual annotation). A second reason is that since the operating environment is not known beforehand, it is impossible to record suitable training material that covers all (still unknown) eventualities that may occur in an unconstrained environment. Consequently, the statistical standard methods are unsuitable for this particular problem.

3. A model of verbal aggression

The starting point for the development of the model for verbal aggression is the assumption that shouts and verbal aggression are behaviorally significant, and that it is therefore likely that relatively low level neural processes are able to focus attention to the target. This implies that simple Auditory Scene Analysis (ASA) can be used to extract an interesting subset of the signal, which can be analyzed further with class-specific knowledge.

This approach has been investigated in the master thesis of Mark Huisman [2] which addresses the estimation of emotional content in speech. Huisman uses Scherer's Component Process theory [3][4] as a theoretical basis.

Scherer describes the reaction of the autonomous nervous system to an emotional stimulus in terms of an activation or deactivation of ergotropic and trophotropic arousal. The first is associated with a stimulation of the sympathic nervous system and leads for example to an increase in heart-rate, blood pressure, transpiration, and adrenalin secretion. An increase in ergotropic arousal results in an orientation reaction due to the stimulus' apparent relevance for the internal goals. This increases awareness of the physical environment and prepares the individual for immediate action. Emotions with a strong ergotropic arousal are panic and anger.

Trophotropic arousal leads to the opposite reaction through a stimulation of the parasympathic nervous system: the organism loses interest in the environment and becomes self-absorbed. Sadness is an example of an emotion associated with trophotropic arousal.

The combination of ergotropic arousal with the muscles and anatomy of the vocal tract results in a number of cues that are in line with the Lombard reflex [5] which occurs while speaking in noisy environments. These comprise increases of the fundamental frequency, the amplitude, and the relative duration of the voiced (periodic) part of speech, in combination with decreases of the importance of the unvoiced fraction and the spectral tilt (which results in spectral "whitening"). Apart from these documented similarities to the Lombard reflex, ergotropic arousal leads to a reduction of voice quality caused by a loss of control over the vocal folds due to over-excitation. Except for the absolute energy, which is strongly dependent on distance, all these cues can be used for verbal aggression detection.

Huisman [2] investigated a large number of numerical cues that can be estimated from the signal with the signal processing approach described in the next section. These cues where correlated with the different emotions of a database of simulated emotions [6]. The three best cues for verbal aggression and panic were fundamental frequency (f_0), the ratio of signal energy below and above 1000 Hz (R_E), and the standard deviation of the energy of the three highest peaks in the spectrum (std E_3). The separation between strong ergotropic emotions and all other emotions in the database in a space spanned by theses three main cues is visualized in Figure 1. The fundamental frequency is the most informative cue.

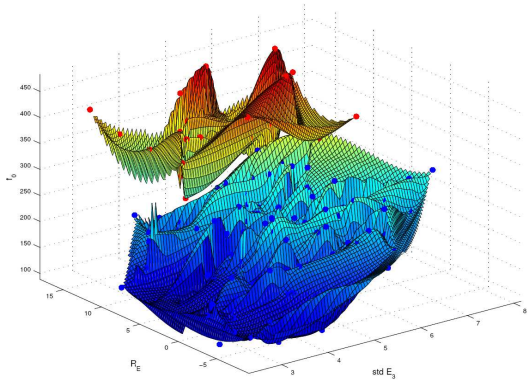

Figure 1 Aggressive emotions are represented by the upper surface

4. System description

The detection system consists of 1) a signal processing stage that simulates some form of auditory attention in the form of foreground/background separation, 2) methods to extract cues for verbal aggression, and 3) a decision mechanism. Together with a description of the physical implementation these are outlined in this section.

Human hearing consists of several stages, the first of which is the conversion of sound waves reaching the ear into a neural signal which can be processed by the brain. In the human inner ear, which is shaped like a snail's shell hence the name cochlea, a structure called the basilar membrane is set in motion by the sound waves reaching it through the outer and middle ear. This basilar membrane has varying mechanical oscillatory properties leading to a nearly exponentially decreasing resonance frequency over the length of the cochlea.

$$f_r(x) = A10^{-ax} - f_0, \qquad \begin{array}{l} A = 17.927 \text{ kHz} \\ a = 60 \text{ m}^{-1} \\ f_0 = 145.4 \text{ Hz} \end{array}$$

(See [7] for a more detailed description of the cochlea.)
Hair cells placed on the basilar membrane convert the motion at each location into nerve activation. This produces a cochleogram [8], a 2D picture of the energy of the sound as a function of time and frequency (as can be seen in Figure 2). This energy was computed using a leaky integration with a time constant of 10 ms.
A sophisticated model of the human cochlea, a transmission-line implementation developed at the Biophysics department in Groningen [9], was used. This ensures minimal information loss in this early stage, which cannot be repaired in later stages.

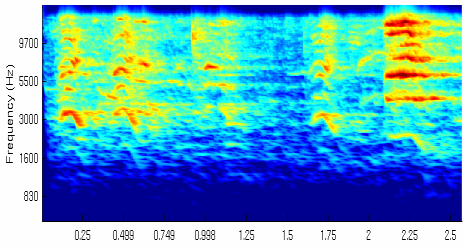

Figure 2: Cochleogram. The last uttering (at t=2.1 s) sounds aggressive**.**

A process located in the first neural relay station - the cochlear nucleus - analyses the temporal dynamics in each frequency channel and performs a first foreground background separation [10]. This attenuates the slowly changing background sounds that do not attract human attention.
This function was implemented in the form of a dynamically adapting background model with a time constant of 10 seconds. If the background model encounters energy more than 6 dB above the current value at any location, this energy is assumed to be caused by a foreground sound, and the corresponding energy is not included in the background model.
So

$$\tau \frac{dE_{bg}(x,t)}{dt} + E_{bg}(x,t) = E'(x,t)$$

where

$$E'(x,t) = E(x,t), \quad \text{if} \quad E(x,t) - E_{bg}(x,t) \le 6\text{dB}$$
$$E_{bg}(x,t), \ \text{if} \quad E(x,t) - E_{bg}(x,t) > 6\text{dB}$$

Very short pulse-like sounds that may occur in microphone signals, are removed with a similar technique as used to compute the background, but now using a time constant of 10 ms. The energy 'filtered' with this time constant, and not 'claimed' by the background model, is the foreground signal (see Figure 3).
This foreground signal is used as a basis for the search for verbal aggression cues. The properties used are: the presence, salience and height of the pitch, the level (compensated for the expected distance between speaker and microphone), the audiblity and three measures for the spectral shape and distortion of the harmonic pattern.

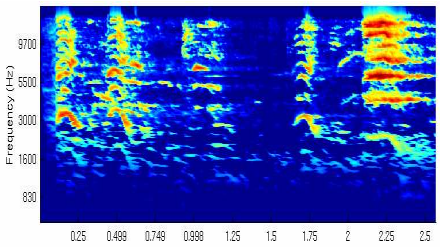

Figure 3 Foreground energy

An efficient implementation of a pitch finding algorithm was used. Herein the average amount of energy of the first n harmonics of a pitch candidate, the sum of the energy at harmonic positions, is compared to the energy at half-harmonic $hh=(n+1/2)*f_0$ positions. So

$$SNR(f_0,t) = \frac{\sum_{x=h_1}^{h_n} E_{fg}(x,t) - \sum_{x=hh_1}^{hh_n} E_{fg}(x,t)}{n}$$

This gives the signal to noise ratio, or salience of the pitch. The frequency value with the highest salience is chosen as the pitch. When the signal is classified as speech-like, it has the right temporal dynamics and a pitch in the human range, voice properties such as the distortion of the voice are determined, which can be associated with stress on the vocal chords. This distortion shows itself as an increase in the high frequency energy, the width and fluctuation of harmonics and in the separation between harmonics. Each sound or voice cue is compared with knowledge about the properties of sounds that attract human attention in general or with values representative of this cue in aggressive sounds. This results in a measure of verbal aggression through the combination of separate cues into an overall pseudo-probability. This is illustrated for the sound used for Figure 2 and Figure **3** in Figure 5.
Figure 4 shows an overview of the whole system.

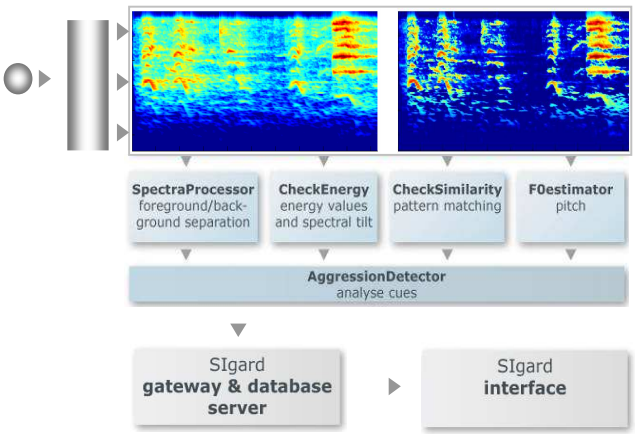

Figure 4 System overview

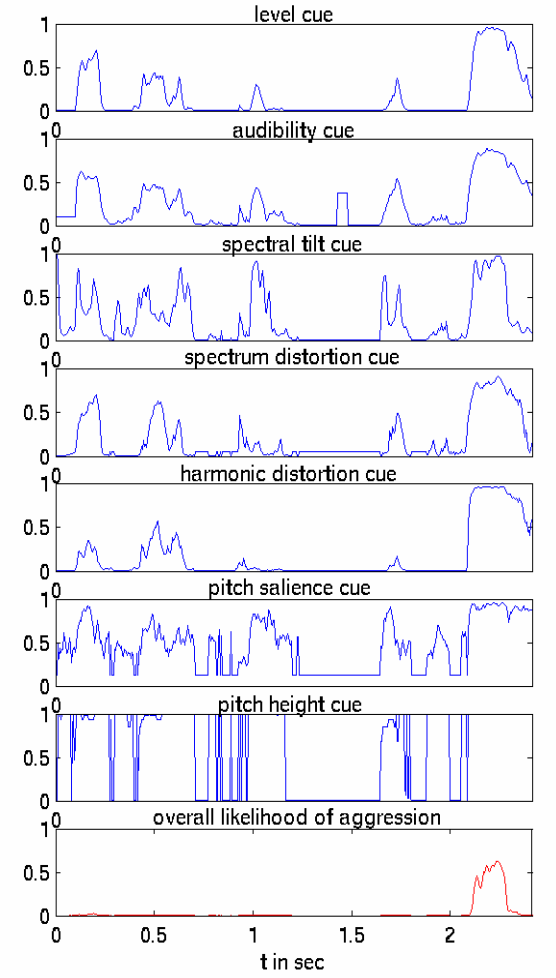

Figure 5 Computation of overall likelihood based on individual cues

If the probability of aggressive shouting is high enough over a sufficiently long interval, an aggression alarm is generated. The thresholds and especially the amount and duration of aggression that needs to be present are parameters that depend on user feedback.

The physical setup consists of a low cost, far field, weather-proofed, microphone with a ~50 dB dynamic range (ClockAudio C007wr). This microphone is connected to specially designed and weatherproof analysis hardware. Detections are routed via IP to a central gateway where they are logged and administrated. A user interface can be configured to give an audio or video alarm on detection and provides an observer with the possibility to add comments and access the logs. Figure 6 shows the user interface.

The varying acoustics of application for example in a narrow street or on a large open square, are measured during the installation phase. In the optimized system, the cochleogram is adapted for the average spectral influence of these acoustics.

Figure 6 User-interface with alarm popup

5. Experimental setup

Measuring the quality of the aggression detection system objectively is not straightforward. During the development of the system there was no scientifically validated database with real-life aggression recordings available on which the system could be validated.

The Cassandra project, a cooperation with the University of Amsterdam, involves the construction of a video and audio database of a range of realistic aggression scenes played by actors in a train station. But although this database contains realistic scenes, the drawback of the use of actors is that they have trained voices, which they always protect during acting. This is a problem, since one of the most important cues in the system relates to the loss of control over the voice. Another problem is that there is no clear, scientifically valid, and objective measure for the level of aggression present in a voice that can be used to form a ground-truth.

These problems prevent a normal scientific validation. But a proper scientific validation is not the best indicator for the successful commercial deployment of new technology: ultimately customer satisfaction in normal operating conditions is much more important. It was therefore decided to let the Groningen police judge the quality of the system according to their needs and expectations, which are in part specific for the local situation. This led to two assessment periods. The first was a three week pilot project at two locations, which took place in the spring of 2005 and the second, which was recently concluded, was the 10 week optimization phase of the official deployment with 11 different microphone positions spread over the inner-city of Groningen.

In both cases the detectors were placed in the center of the pub district. This center services about 40000 visitors each week and has no serious problem with excessive aggression. However, (verbal) aggressive incidents occur with a frequency of about once every week near each of the 7 camera's in this area. During the busiest three evenings and nights of the week, surveillance is intensive (both with cameras and foot patrols). The positioning of the cameras ensures that the probability is low that aggressive incidents will be missed by the human observers. This entails that it is possible to estimate the miss- and false-alarm rate reliably.

The police logged all detected aggressive events during both test periods. At the same time the system produced detections which were stored with a temporal context of 1 minute before and after the detection. All detections were presented to the camera-observers and assigned to different classes, based on their usefulness as supporting evidence for the camera surveillance task.

Several differences exist between the original pilot and the official deployment. In the pilot we used the built-in microphones of the Mobotix M10D camera. However these turned out to have a dynamic range for far-field sounds that was too limited. This resulted in clipping of very loud signals, such as ambulance sirens. Unfortunately, clipping has spectral effects that are quite similar to the loss of voice quality, and therefore the results showed an increased number of false alarms due to sirens.

During the pilot phase the police used four classes for the annotation of the data: essential (indicating aggression and the need to act immediately), useful (indicating aggression but no need to act), justified (no aggression, but the attention of a policeman on the scene would have been grabbed) and false alarms. During the deployment phase only three categories were used: 1) aggression requiring immediate action, 2) aggression not requiring action, or 3) no aggression and therefore a false alarm.

Another difference between the pilot and the deployment was, of course, the larger number of channels and the fact that some of the additional microphones are positioned at locations in which not all acts of verbal aggression will be detected and logged by the police without the support of SIgard. This makes the estimation of the miss rate less reliable, but the police indicated that, although a minor act of verbal aggression may remain unobserved, the large number of foot patrols and the quality of contacts with the public, and the bouncers and staff of pubs makes it unlikely that serious incidents will not be reported. This entails that the miss-rate is fairly accurate when it comes to category 1 detections.

6. Results

During the pilot phase of 18 days a total of 96 detections were produced. The distribution is shown in table 1.

Table 1 Distribution of detections during pilot project.

essential	useful	Justified alarm	false alarm	missed
2	23	44	27	0

During the 10 week optimization phase of the final deployment, the system used weak settings to ensure a large number of recorded false alarms. These recordings were later used to test more restrictive settings. Note that the optimization involves the reduction of the set of accepted events: it is impossible that the optimized settings produce false-alarms not included in the initial database. During the optimization phase the police used the user-interface to annotate all detections. For each of the three optimization rounds, a subset of about 20 detections that were considered 'border cases' was presented in a special session to all camera observers. During this session they could listen to the audio in detail. These detections were re-annotated and used for a next optimization round. The optimized settings were tested against the entire set of initial detections. Table 2 shows the distribution of the entire database of recordings. The first line of results shows the distribution with the permissive setting, the second line the distribution with the optimized settings.

Table 2: Distribution of original detections and the final score with optimized settings.

Aggression + action	Aggression no action	False alarms	Missed alarm
7	34	1359	0
7	22	2	0

The most striking aspect of these results is, of course, the huge effect of optimization on the reduction of the number of false alarms due to the improvement of the aggression model. The final false alarm-rate per channel is exceptionally low given the complexity of this environment. Optimization led also to a reduction of the detections of aggressive shouting which did not require immediate action. In almost all of these cases there was only a single aggressive shout, not followed by further aggression.

An interesting aspect is that the 1359 false alarms in the database contained only 47 detections on sounds other than human shouting. These 47 sounds were almost exclusively sirens, which are designed to attract human attention and have an alarm function much like

aggression. The other 1312 false alarms were cases of people shouting, sometimes very loud and with loss of voice quality due to alcohol consumption, but without being aggressive. These recordings give a good impression of the circumstances under which the detection system operates.

7. Conclusions

The verbal aggression detection system presented in this paper is by necessity an example of a knowledge-driven approach. The model of aggression is based on Component Process theory, a well-established theory of the effects of emotions on speech that leads to predictions that allow the formulation of effective acoustic cues related to psychophysical units like pitch, roughness and timbre. The use of these effective and informative cues allows a decision system with very few parameters in combination with a simple recognition stage. The signal processing is based on models of the auditory periphery in which preservation of information and auditory object tracking are central concepts. Most computational effort is placed on the signal-processing which must ensure that the acoustic cues are as informative as (physically) possible.

Due to the knowledge-based foundation, the classification system is easily extended and requires little retraining when new target sounds are included.

The effect of optimization on the performance is prominent. However, optimization becomes less important as more realistic samples of verbal aggression are recorded at different deployment locations and situations. These are used to determine initial model paramters with a much reduced false alarm rate and speed up the optimization process.

The system described in this paper may be the first detection system for a non-trivial target in an unconstrained environment. The system needs to be fine-tuned for a new environment and for new user demands, but when properly fine-tuned, the system can reach a near human performance (on the recordings). This raises the important scientific question why the system is working so well in complex, uncontrolled social settings that are normally well beyond the scope of standard (sound) classification methods. We believe this is in part due to the fact that traditional classification methods focus on closed domains, while the human perceptive system is optimized for an open domain. Little in the presented approach limits the verbal aggression detection system to a specific domain. In particular the match of explicit properties of a target phenomenon with explicit and meaningful cues estimated from the signal ensures that the system is target specific, but insensitive to the details of general acoustic environments that are unlikely to produce a pattern of cues as seen in verbal aggression.

The results have been judged so impressive that a number of Dutch police departments, the Dutch railway company, and two prisons already consider the system to be indispensable for a modern surveillance system.

Acknowledgements

We would like to thank Jeroen Nederlof and colleagues of the Dutch Railway companies NS and ProRail, the mayor and Municipality of the city of Groningen, the Groningen city police, the Dutch Justice Department, Ruud van Munster and colleagues of TNO, and many others, for their contribution to the development and fine tuning of the SIgard aggression detection system. The support of the Dutch Science Foundation NWO under grant 634.000.432 within the ToKeN2000 program is gratefully acknowledged.

References

[1]. T.C. Andringa and M.E. Niessen, *"Real World Sound Recognition, a Recipe"* (2006)

[2]. M. Huisman, *"Akoestische Effecten van Emoties in Spraak: De Waarneming van Verbale Agressie" ("Acoustic effects of Emotions in Speech: the Perception of Verbal Aggression")*, master thesis Department of AI, University of Groningen (2004).

[3]. K.R. Scherer, *"Vocal communication of emotion: A review of research paradigms"*, Speech Communication 40, pp 227-256 (2003)

[4]. K.R. Scherer, *"Vocal Affect Expression: A review and a model for future research"*, Psychological Bulletin 99(2), pp 143-165 (1986)

[5]. J.C. Junqua, *"The Lombard reflex and its role on human listeners and automatic speech recognizers"*, Journal of the Acoustical Society of America 93(1), 510-524 (1993)

[6]. K. Scherer, H. Wallbott, R. Banse, and H. Ellgring, *"Simulated Emotion database"*, courtesy of The Geneva Emotion Research Group.

[7]. H. Duifhuis, H.W. Hoogstraaten, S.M. van Netten, R.J. Diependaal, and W. Bialek, *"Modelling the cochlear partition with coupled Van der Pol oscillators"* in *Cochlear Mechanisms: Structure, Function and Models*, eds J.W. Wilson and D.T Kemp (Plenum, New York), pp 395-404.

[8]. T.C. Andringa, *"Continuity Preserving Signal Processing"*, PhD-thesis University of Groningen. (2002)

[9]. P.W.J. van Hengel (1996), *"Emissions from cochlear modelling"*, PhD thesis.

[10]. D.A. Godfrey, N.Y.S. Kiang and B.E. Norris, *"Single unit activity in the posteroventral cochlear nucleus of the cat"*, J. Comp. Neurol. 162, pp 247-268 (1975)