Machine Emotional Intelligence: A Novel Method for Analysis of Spoken Affect

Irina Gorodnitsky and Claudia Lainscsek
University of California, San Diego
Department of Cognitive Science
9500 Gilman Dr. La Jolla, CA USA
igorodni@cogsci.ucsd.edu,clainscsek@ucsd.edu

Abstract

Intelligent human-computer interfaces, medical diagnostics, and design of consumer products are just a few of the many applications that can benefit considerably from machine abilities to recognize and adapt to the user emotional state. The paper considers the problem of automatic affect recognition from continuous speech. It describes a new text-independent affect recognition system that is shown to have the capability to provide discrimination of speaker emotional state from acoustic features. The algorithm represents acoustic signals in form of timing sequences. The key feature of the representation is the learning of the timing information directly from the data. In this representation, strong deterministic structures in the data are uncovered. The proposed system exploits the data structure information to perform text-independent affect recognition. The experiments were performed using sentences from a German-language emotional speech corpus [1]. The method is shown to extreact acoustic features that can be used to differentiate neutral, happy, and angry emotions expressed by a speaker.

1. Introduction

Ability to perceive and process emotions is basic to our thinking and day to day functioning [2]. Effective human-computer interactions require that machines possess at least a subset of the emotion processing skills of humans. A computer that recognizes what a person says but ignores how the person says it will appear to be talking but never listening, and is likely to annoy the user.

Affect recognition is useful in a number of other fields outside intelligent machine interfaces, ranging from the medical field, in evaluation of patient emotional states and stress, to ergonomic design, and in assessing user acceptance of ever more complicated consumer technologies by evaluating the level of aggravation arising from user interaction with a product.

This paper is concerned with recognition of emotional expression from speech. Although humans can easily perceive emotions from auditory cues, corresponding machine technology for emotion recognition has been slow to develop. Most current computer technology for affect recognition is based on spectral domain features. Although these have been found very valuable for characterizing speakers and recognizing speech, thus far they have shown only some promise in affect recognition. Currently, no affect recognition systems exists that can be deployed universally, i.e. applied to a range of speakers in a range of operating environments.

The difficulty in recognizing emotions is compounded by their wide range. Not only there are numerous emotional extremes, that sometimes are not even discriminated consistently by human speakers, for example, joy-happiness-elation, but the expression of emotions spans a continuum between these extremes. Emotions that carry a similar connotation, i.e. negative or positive, may have very dissimilar speech features. Anger - sadness - fear are example of emotions that have a negative connotation but are entirely different in their acoustic features. The difficulty is also compounded by the localized and nonstationary nature of emotional expression. A speaker may place the entire emotional expression on one word in the sentence, or spread it over multiple words. Emotional expression can also vary with prosodic content, pitch and pronunciation. Individual speakers can employ different acoustic 'gestures' in their expression of affect.

The paper explores a novel approach for recognizing affect from acoustic-level features extracted from continuous speech. The approach exploits deterministic structures in the acoustic speech signals which it learnes directly from the signals. The structures can be arbitrarily complex, containing any combination of linear (i.e. periodic) and nonlinear (aperiodic) components plus stochastic noise. Being able to analyze a signal that is an arbitrary mixture of lin-

ear and nonlinear parts embedded in noise, without a priori having to establish the nature of the signal, is one powerful feature of this technique.

The structure is recovered from data in the form of a timing sequence. We refer to this representation as the Interval Domain (ID) representation. What is important in this construct is that the parameters of the timing sequence are learned from the data.

The mammalian auditory system is known to encode incoming acoustic signals as timing sequences of neuronal spikes. The mathematical principle for such encoding of a temporal signal as a timing sequence is not known, however. The algorithm in this paper is derived purely by considering projection of noisy data in a feature space described below. It is interesting, however, that this approach results in the same general principle for representation in terms of a timing sequence as that used by the mammalian auditory system.

The method is tested using sample sentences from a German-language emotional speech corpus [1]. A speech feature extracted by the method is shown to correlate with the neutral, happy, and angry emotions expressed by a speaker.

2. Emotional Speech Corpus

The experiments in this paper use seven sentences from the German language emotional speech corpus described in [1]. The entire corpus is comprised of 148 sentences with identical syntactic form (subject-auxiliary-NP-verb), where NP stands for 'the nominal phrase'. The 148 sentences are divided according to their lexical content. The lexical content, neutral, positive, or negative, was determined by having a group of subjects (n=20) rate the sentences. The sentences were recorded while spoken in Standard German by a trained female speaker. Each sentence was recorded and appeared in the database 6 times, using two forms of accentuations (on the NP and on the verb) and three forms of emotional state (happiness, neutral, and cold anger). Thus recorded utterances either matched sentence lexical content or mismatched it. The seven sentences were randomly chosen from the corpus and the six recordings of each of the sentence were provided to us.

One of the objectives in creating this corpus was to examine the connection between affect-dependent acoustic features and the neural responses of listeners, which were monitored using event-related brain potentials (ERPs). One of the study aims was to discriminate the different semantic conditions from listener responses. The use of the two forms of the accentuation were motivated by the hypothesis that the accented syllables are hyper-articulated while unaccented syllables are hypo-articulated. Vocal effort involved in hyper-articulation may produce measurable differences in acoustic features of emotional expression. The match/mismatch between the lexical content and the spoken affect was the other variable condition. The researchers in [1] hypothesized that the mismatch condition would produce a stronger emotive expression.

The objective of the present study is different from the goals in [1]. We are not concerned with discriminating among the different semantic conditions (match versus mismatch, NP versus final verb accent). The aim here is to recognize the expressed affect from spoken sentences. Thus we use only the speech corpus part of the data in our analysis. The different semantic conditions used in the original study yield an interesting dataset for testing affect-recognition in a variety of speech forms.

3. Timing sequence model for affect recognition

The method presented here is derived from the embedding concept of nonlinear dynamics theory [3]. The impetus for its development was the need to analyze real-life data that could originate in unknown environments that are complex, unstructured, and highly noisy.

At the heart of the method is identification of the deterministic structure in data that may be embedded in high amplitude, random noise. The formulation of the problem is straightforward. Given some data, we are interested in identifying the presence of all structures, periodic (linear) as well as aperiodic (nonlinear), they may contain. The problems associated with analysis of stochastically contaminated time series using classical nonlinear dynamics theory has been well described in the past. Casdagli at al. in [4] showed that given even arbitrarily small amounts of noise, some of the degrees of freedom of a system become completely unrecoverable. This means that classical embedding theory cannot be expected to be almost valid when data are almost deterministic. In other words, formal embedding reconstruction is not directly applicable to noisy data.

To deal with this problem several researchers have used projective schemes that identify a manifold in the embedding space (e.g. [5, 6]). The idea is that deviations of a trajectory in the embedding space from a manifold are caused by random noise in the data and the projection onto the manifold filters this noise, thus recovering the deterministic structure buried in data. Such projection techniques have been proposed and demonstrated on a number of signals including speech [5]. The projective techniques used in these works rely on recovery of the manifold from a reconstructed embedding.

The approach taken here is different. One of the immediate restrictions of the classical embedding theory is the fact that information contained in the embedded representation is critically influenced by the choice of embedding param-

eters and in particular the choice of the time delay values. There have been practical examples where the theoretically sufficient dimension can produce less optimal results than using a smaller dimension []. A relevant consensus from the existing works is that to optimize the embedding performance, including minimizing redundancy in the reconstruction, one should consider time delays that are variable in length, not integer multiples of a common lag [7].

We interpret this conclusion in a derivation where we learn the time delay parameters from data. Starting with the classical embedding theory, one can represent the state of an L-dimensional system from its output time series $x(t)$ by constructing an object in a space spanned by the time series and its time-delayed replicas. The object is a *phase trajectory* which can then be written as $\mathbf{x}(t) = \{x(t), x(t-\tau_1), \ldots, x(t-\tau_{D-1})\}$, which is a function of delayed coordinates. Based on the discussion above, we explicitly assume non-uniform delays $\tau_i,\ 1 \leq i \leq D-1$.

The importance of this construct is that it is diffeomorphic to the original phase space for sufficiently large D so that topological properties of the original high-dimensional system are preserved in the embedding under relatively loose restrictions. This means we can extend embedding theory to model the **dynamics** of the system output. Specifically, we express the evolution of the state vector $d\mathbf{x}(t)/dt$ as a function of the phase trajectories, i.e. $d\mathbf{x}(t)/dt = \mathbf{F}[\mathbf{x}(t), \mathbf{x}(t-\tau_1), \ldots]$. This formulation provided a novel data representation strategy. Instead of chosing delays to construct the embedding, we estimate the parameters of the deterministic function $\mathbf{F}$ from the data derivative $d\mathbf{x}(t)/dt$.

We estimate it in the following way. A general non-linear real-valued function can be extressed in a Taylor series expansion of functionals of increasing complexity around some fixed point. When the function $\mathbf{F}[\cdot]$ represents behavior of a dynamical system, that is, a time series model where the input is formed from past inputs $[x(t), x(t-\tau_1), \ldots]$, the expansion becomes a *Volterra series*. We have

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{x}_0 + \sum_{i=0}^{\infty} g_i \mathbf{x}_{\tau_i} + \sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty} g_{i_1,i_2}\mathbf{x}_{\tau_{i_1}}\mathbf{x}_{\tau_{i_2}} + \cdots + \sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty}\cdots\sum_{i_q=0}^{\infty} g_{i_1,i_2,\ldots,i_q}\mathbf{x}_{\tau_{i_1}}\mathbf{x}_{\tau_{i_2}}\ldots\mathbf{x}_{\tau_q} \quad (1)$$

Equation (1) models the linear and non-linear data components as separate model terms. To find a model that is a projection onto a stable manifold, we consider low-order models made of a finite number of leading terms of equation (1). In other words we subselect candidate structures from equation (1) and fit them to data until we identify the smallest best fitting model.

We find the following low-order general structure to work well in many practical applications that we have attempted, including modeling affect expression

$$F(x,t) = \sum_k a_k x_{\tau_i}^l x_{\tau_j}^m, \quad (2)$$

where $x_{\tau_i} = x(t-\tau_i)$ is a delayed data vector with the delay τ_i , $i, j, l, m \in \mathbb{N}_0$ and $\tau_{i,j}$ permiting zero values, i.e. the signal itself.

This idea of restricted complexity of the model, i.e. leaving some of the dynamics unmodeled, plays a key role in the development of the practical algorithm. First of all, it allows us to reduce the computational load in this ill-posed problem to a manageable level so we can solve for the terms of the model. We describe this later when we present the final practical design model. Second, the unmodelled dynamics provides a means to control effect of noise on the estimation, much like the use of regularization in linear estimation.

The model in (2) permits polynomial functions with up to two non-zero delayed data vectors. The linear part of the equation can contain the scaled data itself plus up to two scaled delayed versions of the signal. The non-linear part of the equation permits any number of two term products of data and/or their delayed versions.

The model estimation problem reduces to a two part task: first select an appropriate low order model expansion and then fit the unknown parameters using the derivatives of the measured data. This estimation problem is non-trivial because its highly ill-posed and because the unknown parameters depend non-linearly on the data. We use a genetic algorithms (GA) to perform optimization here.

4. Affect analysis

The analysis in this paper was done using sample sentences from the German-language emotional speech corpus [1] described in Section 2. Each sentence was recorded six times: with three types of affect and two types of accentuations as described in Section 2. Seven sentences, which were numbers 17, 43, 83, 85, 112, 123 in the corpus, were randomly chosen and the six recordings of each of the sentence were provided to us, for a total of 42 records available for analysis.

Bellow is a list of the seven sentences. The set contains 3 sentences rated as having positive lexical content, 3 sentences rated as having negative lexical content and one sentence rated as having neutral content. The number in front of each sentence indicates its position in the database and the sign '+', '-', or '0' indicates the positive, negative, or neutral rating of the sentence.

85 + Sie hat es ans Licht gebracht.
(She brought some facts to light.)

103 + Er hat um ihre Hand angehalten.

(He asked for her hand in marriage.)

112 + Sie hat den Rekord gebrochen.
(She broke the record.)

17 − Sie hat ihn mit der Waffe bedroht.
(She threatened him with the weapon.)

40 − Er hat sie von der Klippe gestoben.
(He pushed her from the cliff.)

83 − Er hat ihn ins Gesicht geschlagen.
(He slapped him in the face.)

123 0 Er hat den Brief geschrieben.
(He wrote the letter.)

Affect-recognition is typically investigated in the lexical context that is either nonspecific or consistent with the exprесed affect. Yet, in practice, many instances can be found where the spoken affect mistmaches the lexical content, sarcastic expression being one example. Affect-lexicon mismatch in the given corpus provide an interesting study case in this respect. The authors in [] hypothesize that a speaker may use a stronger expression of affect in the mismatched condition, thus resulting in stronger acoustic features than in the neutral or matched condition. As shown below, we find strong evidence in our models that supports this.

4.1. Affect recognition model

The affect recognition model was derived from the 42 sentence dataset as follows. The data were resampled from the original 44100Hz down to 8820Hz. Two, three, and four term models from the general model, Eq. (2) were selected and the model parameters were calculated for frames ranging from 20-ms to 100-ms with various overlaps. The model-frame-overlap combination which was consistently selected by our GA algorithm to produce the smallest fit error in all 42 recordings was the two delay 2nd order model

$$\dot{x} = a_1 x_{\tau_1} + a_2 x_{\tau_2} + a_3 x_{\tau_1} x_{\tau_2}. \qquad (3)$$

with 74.3-ms frames and 12-ms updates.

This model-frame-update setting was selected for the subsequent analysis. Parameters for all 42 records and all frames were calculated again using this setting. Analysis of the results is presented next.

4.2. Results from experiments

We summarize the results for the seven sentences each spoken six different ways and present selected results as space allows. Overall, model Eq. (3) was found to successfully discriminate the three utterances of each sentence recorded with different expressed affect: neutral, happiness, and cold anger. This was found for all seven sentences and for both forms of accent. The parameter τ_2 was found to be the main parameter in Eq. (3) that accounted for acoustic features related to the expressed affect.

In the design of a classification system, one generally employs a scoring function made of multiple signal features. Here, since we can identify a single model parameter that clearly correlates with the variations in the affect, it is useful to study τ_2 to gain insight into how affect expression varies with the different utterance conditions that exist in the data.

We first examine how consistent τ_2 responds to affect changes across all 7 sentences. We find the least well defined separation of the three affect conditions when the content sentence is neutral (number $\#123$) when NP was accentuated. NP accentuation is the default accentuation in German for verb-final sentences. Thus this sentence utterance represents the most 'regular' form of speech. The fact that for this sentese the τ_2 response to affect changes is the smallest may support the hypothesis that vocal effort involved in the normal speech is less than in the cases of unusual constructs. Less vocal effort translates into less pronounced affect expression. Nevertheless, even in this case the τ_2 values for the three affect conditions were clearly separable prior to and during the accented part of the sentence. Plots of the raw values of τ_2 for this sentence for both forms of accent are shown in Figure 1. Three piecewise continuous lines can be observed in the plots. In both plots, the upper line corresponds to the neutral expression, the middle line to anger and the bottom line to happiness. In the NP accentuation case, the upper plot, the separation in the τ_2 values was observed most clearly in the first 2/3 of the sentence, and diminished during the unaccented end of the sentence. The middle line (anger) begins close to the upper line (neutral) at the start of the sentence but drops down at approximately the 180th frame. On the other hand, the 'happiness' line stays low throughout the entire first part of the sentence.

In Figure 1(b), the case where the final verb is accentuated, the τ_2 values form distinct piecewise continuous lines. It is obvious that the lines in this case are separated in enough places throughout the sentence to enable one to discriminate between the three affect conditions.

The frames where τ_2 do not form coherent lines, but are randomly distributed, correspond to pauses between within words. The method finds no consistent model in this breaks and much of the signal power is allocated to the misfit error ρ. These frames can be easily filtered out by thresholding across the ρ values.

For comparison we show in Figure 2 the τ_2 values for sentence number 17, which was rated as being negative. In this case τ_2 values for the three affect conditions are clearly separated for both forms of accent. The top to bottom line order corresponds the same emotions as for sentence $\#123$ in Figure 1. In the NP accentuation case, the three affect conditions are separated throughout most of the sentence, coming together only at the point of unaccented verb at the

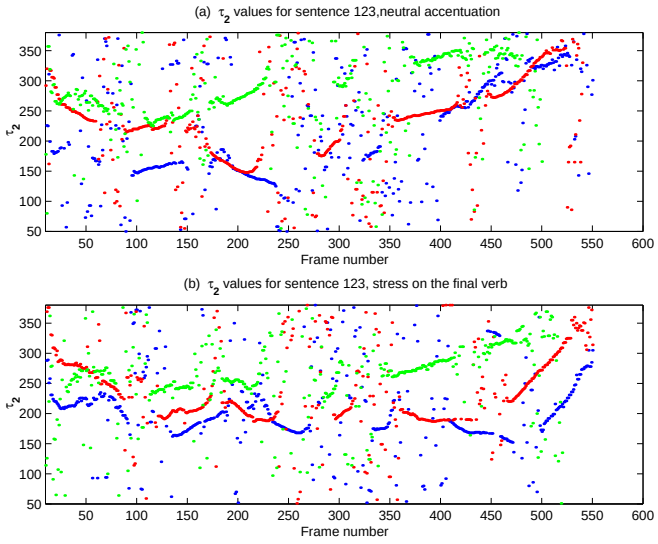

Figure 1. τ_2 values for the three forms of expressed affect in the lexically neutral sentence number 123. (a) Neutral accent. (b) Accent on the final verb.

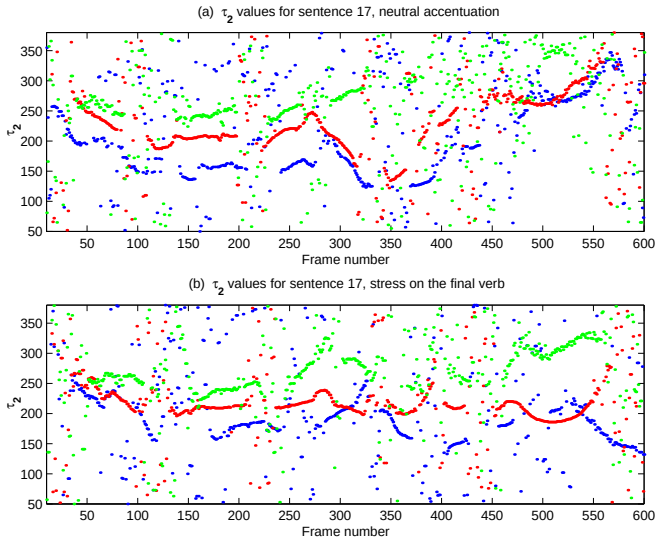

Figure 2. τ_2 values for the three forms of expressed affect in the lexically negative sentence number 17. (a) Neutral accent. (b) Accent on the final verb.

end. In the final verb accentuation case, the τ_2 values are also separated throughout most of the sentence. At the end of the sentence, the τ_2 values for the 'happiness' and 'anger' conditions come together, thus reflecting the accentuation of the final word rather than the difference in affect.

To assay the affect-lexicon independence we compare τ_2 values for each form of affect across the different lexical conditions. Overall, very little difference in τ_2 values was found accross the lexically different sentences for each form of affect expression. Figure 3 shows τ_2 plots for three lexically different sentences (numbers 40, 85, and 123), for two affect conditions, 'happiness' and 'anger', separated by the two forms of accentuation. The only obvious separation in τ_2 values can be seen in frames 325-350 in the top plot of Figure 3. This point is where the main emphasis is placed in the sentence, a point of hyper-articulation.

5. Conclusions

We introduce the concept of timing domain representation, which is implemented by estimating directly from data the delay parameters of a projection onto a manifold in an embedding space. The resulting model estimates deterministic structures in data and the residual stochastic component. Affect recognition power of the model is demonstrated at the acoustic feature level. Seven sentences are selected from a German language emotional speech corpus. The model is shown to find features that distinguish neutral, happy, and angry emotions expressed by a speaker.

The concept of learning embedding parameters from data is in its infancy and it ushers in an entirely new approach. There are a number of areas that require further research, such as design of a classifier based on the presented model, training data requirements, and potential improvement of discrimination by adding conventional speech features to those identified here. More work is planned in this area.

6. Acknowledgements

The authors thank Dr. Gernoit Kubin for providing sentences from the emotional speech corpus.

References

[1] K. Alter, E. Rank, A. Kotz, U. Toepel, M. Besson, A. Schirmer, and A. Friederici. Affective encoding in the speech signal and in event-related brain potentials. *Speech Communication*, 40:61–70, 2003.

[2] A.R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Gosset/Putnam Press, New York, 1994.

[3] Floris Takens. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics*, 898:366, 1981.

[4] M. Casdagli, S. Eubank, J.D. Farmer, and J. Gibson. State space reconstruction in the presence of noise. *Physica D*, 51:52, 1991.

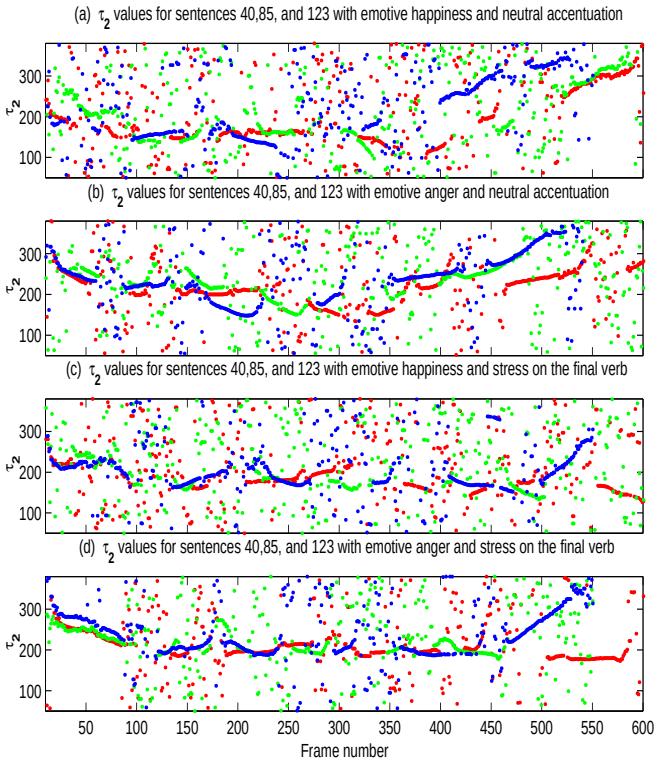

Figure 3. τ_2 **values for three different lexical connotations, sentence numbers 40, 85, and 123 ploted for 'anger' and 'happiness' forms of affect and the two forms of accent. (a) Expression of happiness, neutral accent. (b) Expression of anger, neutral accent. (c) Expression of happiness, accent on the final verb. (d) Expression of anger, accent on the final verb.**

[5] R. Hegger, H. Kantz, and L. Matassini. Denoising human speech signals using chaoslike features. *Physical Review Letters*, 84:3197–200, 2000.

[6] L. Matassini, H. Kantz, J. Holyst, and R. Hegger. Optimizing of recurrence plots for noise reduction. *Physical Review E*, 65, 2002.

[7] P. Grassberger, T. Schreiber, and C. Schaffrath. Nonlinear time sequence analysis. *Int. J. Bifurcation Chaos*, 1:521, 1991.