Visualizing Dynamic Features of Expressions in Speech

Tal Sobol Shikler and Peter Robinson

Computer Laboratory
University of Cambridge, UK
ts313@cam.ac.uk

Abstract

In this paper we examine some of the issues involved in analyzing the expression of emotions, mental states and attitudes in speech. We investigate timing issues that may affect the structure of future automated inference systems. These timing issues include tracking transitions and changes over time within expressions, existence of thresholds between expressions, and the necessity of tracking time-frequency features within an utterance. We also suggest an outline for a system that includes these features.

1. Introduction

Human-computer interaction and human-human communication via computer interfaces still lack the basic means of recognizing and responding to non-verbal cues of attitudes, emotions and mental states, that we take for granted in human communication. This problem is more acute in speech interfaces. In these systems speech is used to convey commands and data, while natural behavior also uses speech for thinking out loud, expressions of frustration, misunderstanding, discomfort, and more. Most of these functions relate to subtle expressions and to nuances of expressions, some of them are obvious only in speech. The inference of such expressions may contribute not only to better dictation and other speech-dependent computer interaction systems, but also to the development of a large variety of aids such as feedback systems that allow people to learn, improve, or practice communication skills. Other applications include remote diagnosis of extreme situations and mental states. In addition, the understanding and automation of affect recognition can contribute also to the generation of more natural synthesis.

Most research in this field is analysis of basic and extreme emotions like joy, anger, sadness, fear, disgust and surprise [1-3]. A few other mental states like stress, frustration and depression have also been investigated in the context of automated systems [4-7]. Several projects have tried to map additional emotions according to their prosody and articulation characteristics [8-10]. Vocal analysis has been used also for identifying insurance fraud.

However, expressions in general, mental states, behavioral patterns, attitudes and personality traits have not been thoroughly investigated. Nor have the timing characteristics of continuous expressive speech been investigated.

We have shown that subtle expressions, spontaneously evoked by human computer interaction tasks, such as uncertainty and enthusiasm, are separable by analysis of non-verbal speech cues [11]. However, we have pointed out that a thorough analysis requires inclusion of timing considerations.

In this work we examine and visualize some of the timing characteristics of naturally evoked subtle expressions. We demonstrate the changes among sentences that relate to the same expression over time, we show the differences among sentences that convey different expressions, and we draw conclusions concerning the structure of automatic inference systems of such expressions.

In Section 2 we presents some of the complexity of dynamic analysis, accompanied with experimental results, in Section 3, we present an outline for a solution, and Section 4 concludes with a discussion and suggestions for future work.

2. Timing of expressions

Non-verbal speech cues can be measured only during speech segments. Most people do not speak continuously and endlessly, but rather in time segments, whose duration and timing are set according to the personality and the nature of the interaction. These speech segments are discrete in time, and usually not equally spaced. Within each speech segment the speech can be viewed as continuous, but even within these semi-continuous speech segments, there are shorter segments or speech units. Speech rate, which measures the distances among speech sub-units and their lengths, is among the parameters that identify the expression. A certain expression can be continuous among several speech segments, or change gradually among them, while there may be speech segments of short-term expressions that interfere with this continuous and leading expression, and even cause minor changes among them. An analysis of continuous interaction should integrate the analysis of the current speech segment, and all its variations, with the analysis of the transitions among speech segments.

2.1. Continuity and thresholds among utterances

A possible complication for the analysis of subtle expressions is that it cannot be done by using blind clustering methods, which are based on distances among samples, because sometimes samples from different groups, labeled as different expressions, are closer to each other then samples from different groups. One reason is the gradual transitions between expressions and within nuances of expressions over time. The second is that different expressions are noticed if a certain threshold has been passed. An example for transitions among expression nuances in the same expression class, and for classification which cannot be supported by blind clustering, is shown in Figure 1.

The graph illustrates the vocal features that separate the expressions of uncertain, cheered, and enthusiastic. Each point in the graph indicates a different speech segment. Segments are different sentences. The sentences were taken from the Doors database [11], in which the same sentence was uttered a hundred times, during a computer gambling game.

The sentences were uttered after the gain, either profit or loss, of the last choice made by the user, had been seen. Each

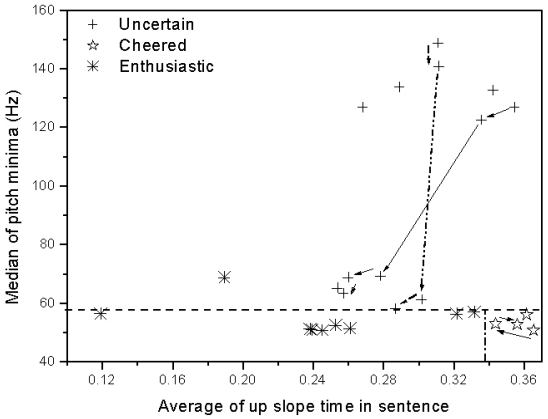

Figure 1: Transitions rules among the expressions of uncertainty, cheered, and enthusiastic, found by a data-mining algorithm. The arrows indicate transitions among consecutive speech segments. The thresholds for transitions between expressions are marked by dashed partitioning lines. The features that distinguish between the groups are the median of local minimum pitch values, and the relative time of pitch rising. The samples were taken from an utterance of a male speaker in the Doors database.

sentence was uttered separately, while in between, another event had happened and another sentence had been uttered.

The text in all the sentences is identical: *sgor de-le-t*, which means close door, in Hebrew. The text repetition and the context neutrality allow specific examination of the changes caused by expressions of different mental states, attitudes and emotions.

The features for classification are the median of minimum pitch values, and the relative time of rising pitch. The rules for classification were derived with the c4.5 algorithm, using the Weka data-mining tool [12]. These rules and features were extracted out of 80 non-verbal vocal cues (prosody) features, which were extracted for each processed sentence, or speech segment. These features included statistical features of the fundamental frequency, the total energy, and the energy in different frequency bands, as described in [11]. It can be seen that some of the samples that were labeled manually as uncertain, are closer to the segments labeled as cheered or enthusiastic, more than to the other samples of uncertainty. These relative locations make blind clustering inefficient in this case. The transitions among the expression behave like passing a threshold. The arrows indicate transitions among speech segments that followed each other in time, in the space defined by the significant features, within the same perceived expression. The time gap between adjacent samples was in the range of 6-9 sec. It can be seen that different speech segments have the same expression, with a tendency towards continuous transitions. Continuous analysis can reveal tendencies in the interaction that the system would like to change in an early stage, such as boredom or frustration. Taking the segments separately could also lead to wrong classification, but following the transition can correct it.

Although the statistical features allow us to distinguish among sentences that convey certain expressions, and to track tendencies and transients, they fail to give us the means of producing the same effects in synthesized speech.

Therefore, we decided to investigate the time-frequency characteristics of each speech segment and the differences among consecutive segments.

2.2. Time frequency variability within an utterance

Another way of visualizing the differences among expressions in speech is to use spectrograms. Although the fundamental frequency, which reveals the intonation, is not very clear in this representation, it reveals other parameters of time and frequency variability which cannot be as clearly observed using other processing methods. An example is shown in Figure 2.

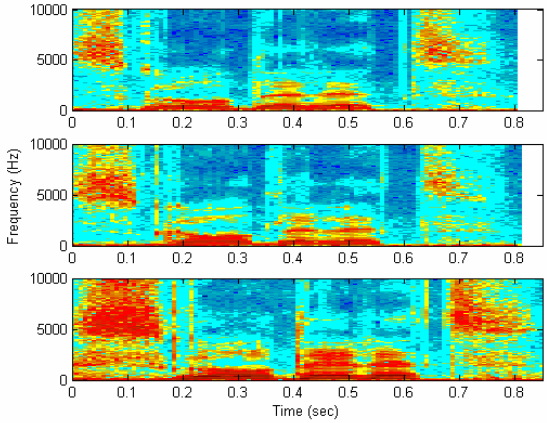

Figure 2: Spectrograms of three consecutive utterances of the sentence *sgor de-le-t* (close door), uttered by the same person, with expression of uncertainty. The red (dark) colors indicate higher energy.

The three spectrograms are of consecutive sentences of *sgor de-le-t*, as in Figure 1, uttered by the same speaker, with the expression labeled as *uncertainty*. The spectrograms reveal different parameters that may signify the gradual change among the instances, such as the gradual change in the duration of voiced parts, the distances between consecutive speech parts, the total energy, and its distribution in the frequency domain.

Figure 3, on the other hand, represents the same sentence, by the same speaker, but with various subtle expressions that were evoked during the interaction. The first two are of the same labeled expression of *uncertainty*, the third relates to an expression of *testing*, the fourth is *cheered* and the last one relates to *down*, or withdrawn. It can be seen that although the first two utterances are not identical, they are very similar to each other, especially when compared to the rest of he samples.

The significant parameters are the sentence length, the length of each uttered part, and the distances among these parts. The intensity or the amount of energy in each uttered part, and the distribution of the energy along the frequencies' spectrum, in various points in time. Whereas *cheered* is signified by a short sentence, short units and very distinct energy bands (red). *Testing* reveals a broken third part and a lengthy and blurred ending part, all the parts of this

expression are intensified, especially the last part. *Down*, as may be expected, is pronounced in a longer sentence, with longer uttered parts and longer pauses, the uttered parts contain less energy, and the third uttered part nearly disappears.

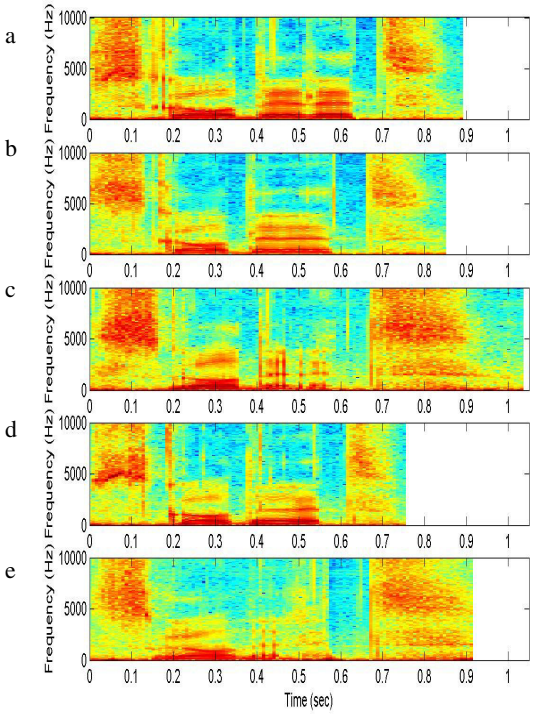

Figure 3: Spectrograms of the sentence *sgor de-le-t* (close door), uttered by the same person, with naturally evoked expressions of uncertainty (a,b), testing (c), cheered (d), and down (e).

One of the problems that the spectrograms reveal is that statistics and averaging over the whole sentence are not good enough for revealing the nuances of subtle expressions. The next question is how spectrograms reveal expression cues in shorter and longer utterances.

Figure 4 shows four strips of spectrograms of the word *lo*, which means *no*, in Hebrew. This word has been chosen because it is very short and very significant – it can be uttered with many different expressions like questions, refusal, scorn, astonishment, and more. Each strip includes 8-11 utterances of *lo*, uttered by the same speaker; strips a, and b by male speakers, while strips c and d by females. It can be seen that the inter-speaker variability is very significant. The features that should be examined change from one speaker to another, and in particular the frequency range that should be examined. While most of the energy of the utterances of the first and fourth speakers is located in the range of 0-5KHz, the frequency range of the other speakers reaches 10KHz, (one of them is a male speaker). The third speaker has a frequency band at 5 KHz, in which there is no energy at all. The voice quality, as revealed by the distribution of energy along the frequency axis, changes among speakers and among expressions of the same speaker. The length of the utterance changes, although here there are no unvoiced parts.

2.3. Conclusions

The complexity of analysis and automatic inference of real, continuous interactions is demonstrated in the presented examples. In some cases, a single syllable can contain all the information, while in others, transitions among consecutive utterances reveal the underlying expression, or follow changes such as building of frustration. The complexity is intensified by the fact that many of the utterances include mixture of expressions, related to personality traits, attitudes to the situation, long term mental states, short-term responses to events, and more. Some or all of these parameters may be represented within a single utterance, or a stream of utterances. Therefore an inference system should be able to support both stand-alone expressions, and transitions. It should have memory, use thresholds, and allow continuity.

We have observed features that should be included in automatic inference systems. In addition to the statistical features over the whole utterance, which were explored in

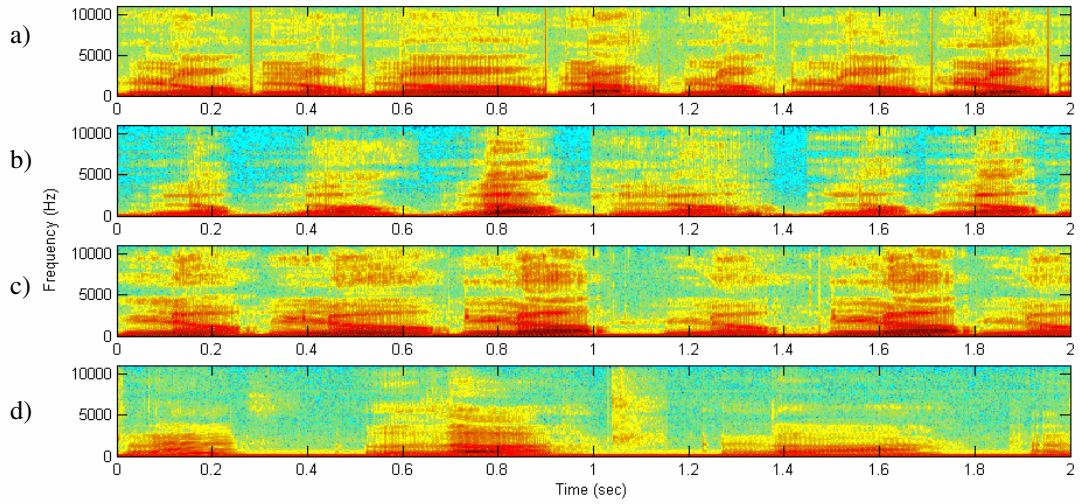

Figure 4: Spectrograms of the word/sentence "lo" (no), uttered with different expressions by two male speakers (a,b), and two females (c,d).

many work [2,13], which were also used for the data mining presented in this paper. Many of the features are time dependent within an utterance. Therefore statistics over a whole utterance is not always good enough. Time related features, and time-frequency related features should be included in the inference. There have been attempts to analyze the behavior of these features over time [4]. However, some of these features do not easily formalize into strict statistical features, with firm boundaries, but are rather fuzzier in nature, and should get a different representation. We believe that some of these features may be extracted using image analysis.

3. Solution

The proposed solution for an inference system should support both stand alone expressions and utterances, and dynamic changes and transitions. If we consider each expression as a state, there are transition rules among states. Some of the states can be reached independently, without regard to the previous state, by the rule itself, while others can be reached only through gradual transitions. The change rules from the current state can lead to several other states, the transition itself will occur only if a threshold is passed. The certainty or the purity of an expression is achieved gradually, when more supportive utterances and rules increase the confidence. Subtle interim expressions can be viewed either as expressions on their own accord, as a variation to a transition rule, or as events, kept in memory in order to support future decisions. The transition rules depend on the current state, and previous decisions. They should include the most significant features, extracted from the current speech signal, either in a stand alone manner or as the gradient among the last few utterances.

4. Discussion

We have presented some of the dynamic characteristics of expressions and subtle expressions of emotions, mental states, attitudes, and display rules, in natural speech.

The nature of these expressions, and their manifestation in speech signals, include gradual changes within expressions, and thresholds for transition among distinguishable expressions. We have shown that traditional blind clustering techniques are not always applicable. We also presented examples of features that can distinguish among expressions, but cannot be described using statistics over the whole sentence. We presented the differences among subtle expressions in short sentences and in very short utterances, with expressions that should be distinguishable in a stand alone manner and as part of transitions. We mentioned the differences among speakers. Although all these parameters are known theoretically, they have not been addressed in the context of dynamic analysis and automatic inference of subtle expressions, in natural and continuous speech. We suggest an outline for a model, or an algorithm that addresses these issues. We are examining ways to apply this model.

Further consideration should be given to the problem of labeling subtle and mixed expressions. This problem affects both the organization of reference data, and the learning process. Image processing and other method to extract and represent the time-frequency features should be explored, and ways to represents transitions among expressions in time should be investigated.

The inference does not have to depend only on speech, in many cases additional data can be added to a system to support the decisions. The additional data can include other human communications cue, such as facial expressions, physiological measurements, like heart bit rate, and reports of events and context. All these parameters are additional to the features which can be extracted from the speech itself.

5. References

[1] Lisetti C.L. and Schiano D.J., "Automatic facial expression interpretation: Where human-computer interaction, artificial intelligence and cognitive sciences intersect", *Pragmatics & cognition*, 8(1), 2000.

[2] Oudeyer P.Y., "The production and recognition of emotions in speech: features and algorithms", *International Journal of Human Computer Interaction*, 59(1-2): 157-183, 2003.

[3] Yacoob Y. and Davis L.S., "Recognizing human facial expressions", *Image Understanding Workshop. Proceedings*. San Francisco, CA, USA, 1994.

[4] Cohn J.F. and Katz G.S., "Bimodal expression of emotion by face and voice. Workshop on Face / Gesture Recognition and Their Applications", *The Sixth ACM International Multimedia Conference*, UK, 1998.

[5] Fernandez R. and Picard R.W., "Modeling drivers' speech under stress", *Speech Communication*, 40: 145-159, 2003.

[6] Guojun Z., Hansen J.H.L. and Kaiser J.F., "Classification of speech under stress based on features derived from the nonlinear Teager energy operator" *Proceedings of the ICASSP '98*, New York, USA, 1998.

[7] Klein J., Moon Y. and Picard R.W., "This computer responds to user frustration: theory, design, and results", *Interacting with Computers*, 14(2): 119-140, 2002.

[8] Cornelius R. and Cowie R., "Describing the Emotional States that are Expressed in Speech", *Speech Communication*, 59, 2003.

[9] Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W. and Taylor J.G., "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, 18(1): 32-80, 2001.

[10] Scherer K.R., "Emotion effects on voice and speech: Paradigms and approaches to evaluation", *ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[11] Sobol Shikler T. and Robinson P., "Recognizing Expression in Speech for Human Computer Interaction", *Designing a more inclusive world*, Keates, S., Clarkson, J., Langdon, P., Robinson, P., (Eds.), Springer-Verlag, UK, 2004.

[12] Witten I.H., Frank E., *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.

[13] Dellaert F., Polzin T.h. and Waibel A., Recognizing emotions in speech. *ICSLP 96*, 1996.