

Experiments in Emotional Speech

Julia Hirschberg, Jackson Liscombe, and Jennifer Venditti
Columbia University
{julia,jaxin,jvv}@cs.columbia.edu

February 18, 2003

0.1 Introduction

Speech is a rich source of information, not only about what a speaker says, but also about what the speaker's attitude is toward the listener and toward the topic under discussion — as well as the speaker's own current state of mind. Until recently, most research on spoken language systems has focused on propositional content: what words is the speaker producing? Currently there is considerable interest in going beyond mere words to discover the semantic content of utterances. However, we believe it is important to go beyond mere semantic content, in order to fully interpret what human listeners infer from listening to other humans.

In this paper we present results from some recent and ongoing experiments in the study of emotional speech. In Section 1 we discuss previous research in this area, and in Section 2 we describe a recent and several planned experiments addressing important methodological issues in the study of emotional speech. We conclude in Section 4 with remarks on the ultimate application of results from these experiments to the automatic identification of emotion in speech.¹

1 Previous Research

In recent years there has been considerable research, both theoretical and empirical, on the perception and production of emotional speech. Theoretical work of psychologists and speech scientists has focussed on the development of general frameworks within which emotions can be categorized, (Cornelius, 2000; Cowie, 2000; Gussenhoven, 2002; Pollermann, 2002; Scherer, 2000). In this, researchers attempt to define 'emotion' as a concept as well as identifying theoretical constructs that individual emotions participate in to varying extents, to account for

their similarities and differences. Experimental work has sometimes tested these theoretical proposals but has more often attempted to identify, independent of theory, some set of features that reliably distinguishes one emotion from others in forced choice tests (Cowie and Douglas-Cowie, 1996; Kienast and Sendlmeier, 2000; Mozziconacci and Hermes, 1999; Pereira, 2000; Schröder, 2001; Yuan, Shen, and Chen, 2002). Acoustic and prosodic features such as intensity, duration, speaking rate, spectral balance, phonation, articulation, fundamental frequency (F0) range and mean, and overall intonational contour are then calculated for utterances labeled with the same emotional state and descriptive statistics obtained by way of characterizing such states. Results have been promising for some emotions and some languages.

However, many of the most reliable features identified in this way require laborious hand-labeling, and thus are of little practical use for computational modeling. And while empirical studies report subsets of features significantly associated with different emotional states and confusion matrices for subjects judgments, little attention is paid to relationships among various cues to emotional state: e.g., which are necessary and which are sufficient? Which are redundant? Thus we do **not** have a good understanding of which emotions are perceived as similar and what the underlying acoustic, prosodic, lexical or contextual explanation might be. Most studies have indeed found that some (varying) subset of emotions prove difficult for subjects to distinguish reliably, such as, e.g., *anger* and "frustration" or "happiness" and "engagement", and all laboratory studies suffer from the artificiality of the task at hand. Since it is difficult to convey a clear description of the emotions to be labeled, some studies have included subjects' confidence ratings with each judgment but none to date has permitted them to assign multiple emotional labels. And no studies have discovered a reliable method for eliciting objective data that might mediate the 'noise' widely recognized in these difficult subjective judgments.

¹Thanks to Dan Jurafsky, Brian Pellom, Liz Shriberg, and Andreas Stolcke for useful discussions.

More focussed research has been done by speech technologists, seeking to identify useful parameters to vary in speech synthesis (Schröder, 2001) or to identify during speech recognition/understanding, especially in spoken dialogue systems (Ang et al., 2002; Litman, Hirschberg, and Swerts, 2001; Batliner et al., 2000). Experimentation in the former domain has generally followed the same lines as described above, in which actors read utterances trying to convey particular emotions, those utterances are classified by listeners, and utterances which score high for particular emotions are analyzed for their acoustic/prosodic features. These parameters are then varied (with more or less success and considerable human intervention) in a given text-to-speech system, to convey the desired emotion, and listeners are asked to rate the synthetic speech (Cahn, 1988; Burkhardt and Sendmeier, 2000; Murray et al., 2000). For synthesizers providing less direct control over acoustic parameters, emotion experiments have been done in which 'emotional' inventories are recorded and utterances produced from them judged by listeners as to affect (Bulut, Narayanan, and Syrdal, 2002). Since the goal of such studies is to produce emotional speech automatically, most are confined to investigating features such as F0, timing and loudness, which can be manipulated in systems systematically.

Promising work has been done recently in emotion detection in meetings, voicemail, and in spoken dialogue systems, especially for English and German (Ang et al., 2002; Litman, Hirschberg, and Swerts, 2001; Batliner et al., 2000). These corpus-based studies have addressed the problem of emotion detection in natural or elicited speech, attempting to detect emotions such as anger and frustration with system problems in system-user interactions or urgency in voicemail by hand classifying or rating instances, extracting acoustic and prosodic features, such as duration, pitch, and energy as well as lexical cues, and employing machine learning techniques to develop predictive models. Success rates have ranged from 60-80%, depending upon the distinction attempted, with frustration and anger detectable in German with about 60% success (Batliner et al., 2000; Huber et al., 2000) and in English with 60-80% accuracy (Ang et al., 2002; Lee and Narayanan, 2002), on different corpora and with differences in definition of target emotion. Significant improvement was found also in voicemail ranking by urgency or personal nature using a variety of metrics.

Despite these promising beginnings, several critical and related barriers stand in the way of our understanding of emotional speech: a) we lack the ability to elicit reliable human judgments classifying particular speech tokens with particular emotion labels, whether in the labora-

tory or in labeling corpora; b) we thus lack large reliably-labeled corpora on which to train methods to detect emotion in speech; and c) we thus are unable to explore the full set of features, acoustic, lexical and contextual, which may prove useful in identifying different types of human emotion in speech. In this paper we focus on an ongoing project addressing some methodological barriers to the study of emotional speech.

2 Methodological Issues in Understanding Emotional Speech

A critical problem in past studies of emotional speech detection, whether in laboratory or corpus-based experiments, is how to elicit from human subjects or labelers in the lab, reliable judgments of emotional state from spoken inputs. Without reliable judgments, little can be learned about the features that convey speaker affect and thus useful computational models of emotion cannot be constructed. In most laboratory studies, simple classification judgments are solicited for isolated utterances usually performed by an actor (e.g., "Classify this utterance as **either** *angry* or *sad* or *frustrated* or ..."). But utterances may convey multiple emotional messages to hearers, all of them part of the speaker's intent. Also, while the hope is that listeners share some core notion of basic emotions, it is not always clear that they operationalize their labeling task similarly: internal subjective impressions must be translated into a simple decision. And some emotions have been difficult for subjects to discriminate between, such as, e.g., *anger* and *frustration* or *happiness* and *engagement*. There is considerable disagreement in labeling tasks as well: (Lee and Narayanan, 2002)'s corpus of call center interactions was labeled for 'negative' emotion by two labelers, who only agreed in 65% of cases. To make the task easier, some studies have allowed subjects to provide confidence ratings with each judgment, to ease the burden of decision. However, to date, none has permitted them to assign multiple labels to a single utterance. And no studies have discovered a reliable method for eliciting *objective* data that might mediate the 'noise' widely recognized in these difficult subjective judgments. To address these problems in judgment elicitation, we propose two new paradigms for obtaining human judgments of emotional speech: a) eliciting multiple emotion rankings of emotional speech tokens from listeners, and b) obtaining objective reactions as well as subjective judgments of emotional tokens via a series of eye-tracking experiments.

2.1 Rating Utterances on Multiple Emotion Scales

To explore the first paradigm, we have recently conducted a web-based pilot study to discover whether it is possible to obtain multiple-emotion ratings of emotional speech tokens. We have preliminary results not only validating our hypothesis but also pointing us to further investigations of how perception of particular emotions is correlated and some hypotheses about some acoustic cues which may explain some of these relationships.

For this pilot, we selected tokens from the LDC Emotional Prosody Speech and Transcripts corpus (<http://www ldc.upenn.edu/Catalog/LDC2002S28.html>), and prepared a web-based experiment, in which subjects were asked to rank each utterance on multiple scales. The Emotional Prosody corpus contains recordings of 8 professional actors (5 female, 3 male) reading short (4-syllables each) dates and numbers (e.g., “two thousand four”) in 15 distinct emotional categories: *disgust, panic, anxiety, hot anger, cold anger, despair, sadness, elation, happy, interest, boredom, shame, pride, contempt, and neutral*. Actors were used because they are trained to produce a range of emotions in a convincing manner. The actors were given descriptions of each emotion, along with several examples of situations in which that emotion would be appropriate. They were allowed to repeat each phrase as many times as they wanted, resulting in numerous tokens of each phrase in each emotional category. For our experiment, however, we modified the set of categories to be rated to represent emotions we felt were particularly important to the corpora we will ultimately examine.

Positive emotions: *confident, encouraging, friendly, happy, interested*

Negative emotions: *angry, anxious, bored, frustrated, sad*

One token representing each category plus *neutral* was selected from each of 4 actors from the corpus (MM and GG (female), CC and CL (male)), resulting in a total of 44 utterances. Selection was determined by listening to all of the LDC tokens and finding convincing exemplars matching each of our emotion categories. In addition, 3 more tokens were chosen from 3 other actors to use in practice trials. This resulted in 47 utterances total for the survey.

Subjects participated in the survey over the internet. After answering introductory questions about their language background and hearing abilities, subjects were

Emotion Recognition Survey: Sound File 1 of 47

	not at all	a little	somewhat	quite	extremely
How frustrated does this person sound?					
How confident does this person sound?					
How interested does this person sound?					
How sad does this person sound?					
How happy does this person sound?					
How friendly does this person sound?					
How angry does this person sound?					
How anxious does this person sound?					
How bored does this person sound?					
How encouraging does this person sound?					

Play Next Item

User ID: 8668462401

Having trouble with the survey? Please email the webmaster and include your user ID listed above.

Figure 1: Example response page from web-based perception experiment.

given written instructions describing the procedure. Subjects were asked to rate each token (which played out loud over headphones or speakers) on each of 10 emotional scales (see above, a ‘neutral’ scale was not included). For each emotion, subjects were asked *How X does this person sound?*. Subject responses could include: *not at all, a little, somewhat, quite, or extremely*. At the start of the experiment, subjects were presented with the 3 practice stimuli in fixed order. Then the remaining 44 test stimuli were presented one by one in random order. For each stimulus trial, a grid of blank radio-buttons appeared, as depicted in Figure 1. The sound file for that trial played repeatedly every two seconds until the subject selected one response for each emotional scale. Subjects were not allowed to skip any scales.

The order in which the emotional scales were presented was rotated among subjects. Two randomized orders and their reverse orders were used, resulting in 4 distinct orders. In addition, the first emotion displayed in the response grid was varied across trials such that a given emotion which appeared first in one trial would appear second in the next trial, in a cyclical manner. In other words, for a given subject, the order of the emotion categories was fixed, but the order in which the categories were displayed was rotated from one trial to the next.

Forty-seven of the 189 subjects who began the study completed it, for a completion rate of 24.9%. On average, those who aborted the survey did so early, after item 3. We

have excluded from our analysis subjects who reported hearing impairment or who were not native speakers of Standard American English.

Of the 40 subjects thus analyzed, 17 were female and 23 male. All subjects were 18 or older, with a fairly even distribution among age groups. Mean time spent on the survey was 48 minutes and median time 32. A correlation matrix for subject ratings of each token on each emotional 'scale' is presented in Table 1, where correlations were calculated for each pair of emotions from each subject's rating of each utterance on those scales ($n=1760$) ($r > = .195$ is significant at the .05 level):

Frustration patterns as we might expect, with strong positive correlation only with *angry*, and strong negative correlations with *encouraging*, *happy*, and *friendly*. There are also intuitively plausible correlations between *friendly* and *encouraging*, *happy*, *interested*, and *confident*. Note also that *bored* is positively correlated with *sad* and negatively with *happy*. Since TTS systems are routinely criticized as sounding *bored*, do they also sound unhappy? *Sad* is negatively correlated with *confident*, *friendly*, *encouraging*, *interested*, and, of course, *happy*. It is interesting that the speaker's own personal state, *sad* or *happy*, seems to carry over into more other-directed states, such as *encouraging* and *interested*. And one emotion — *anxious* — is not correlated with any other.

To identify which acoustic and prosodic features identify an utterance as conveying one or more of these emotions, we also conducted a preliminary analysis of some simple acoustic features, normalized for speaker characteristics, including mean fundamental frequency (F0) (meanF0) and RMS (meanRMS) over the utterance, F0 and RMS maxima within the nuclear stressed syllable (F0 event, maxRMS), and syllables per second (rate). These features roughly capture perceived pitch range, loudness, and speaking rate. We then calculated means for each token and correlated each with mean subject rating for that token along each emotional scale. For sample size 44 ($df=2$), r of .304 and above would be significant at the .05 level. Table 2 shows results from this analysis:

In this table we see immediately that our simple features appear to be correlated significantly only with *bored* (-meanF0, -F0event, -meanRMS, -maxRMS), *confident* (+meanF0, +F0event, +meanRMS), *encouraging* (+meanF0, +F0event), *friendly* (+meanF0, +F0event), *happy* (+meanF0, +F0event), *interested* (+meanF0, +F0event, +maxRMS), and *sad* (-meanF0). So, utterances with lower pitch and amplitude are rated as *bored*, while those with higher pitch and amplitude are rated *confident* and *interested*. Higher pitch appears also to lead to higher ratings of utterances as *encouraging* or *friendly* or *happy*,

while lower pitch characterizes *sad* utterances. Note that utterances that exhibit higher pitch features (*confident*, *encouraging*, *friendly*, *happy*, and *interested*) are also highly correlated in the speaker ratings in Table 1. It thus appears that acoustic cues will provide excellent sources of information about how emotions are perceptually related and good cues to recognizing them.

3 Future Experiments: Eye Tracking and Emotional Speech

Our next set of methodological experiments will supplement subjective judgments of emotional state with objective data obtained from tracking subjects' gaze during subjective decision-making. Our laboratory experiments will include a series of eye-tracking experiments designed to examine listener judgments of emotional speech. We choose eye-tracking as a means to obtain an objective measure of listener preferences, without explicitly asking for overt judgments about emotion. In the eye-tracking paradigm, listeners view pictures on a computer screen while they hear auditory stimuli. Eye movements and fixations are monitored by an ISCAN Inc. remote infrared-reflection eye-tracking device which is positioned on the table-top in front of them. (This procedure poses minimal risks to human subjects, and has been approved for use in psycholinguistic research by the Columbia University Institutional Review Board (protocol #02/03-937A).) Eye-tracking has become a popular methodology for research in spoken language processing, since it allows experimenters to monitor (visual) attention to entities without interrupting the speech stream (unlike the gating paradigm, for example), and because eye movements have been found to be closely time-locked to the auditory input (See (Tanenhaus et al., 1995) for a brief introduction to the eye-tracking paradigm).

Experiment 1: Associating emotional speech with emotional faces: The first experiment is designed to provide an objective measure of perceived emotional state without explicitly eliciting listener judgments. Listeners will view a set of pictures of faces exhibiting various emotional states (e.g., anger, sadness, happiness, etc.) presented on the computer screen. The pictures will be selected from the set of 110 photos published by Ekman & Friesen (Ekman and Friesen, 1976), which have been independently rated for emotional state. While viewing the pictures, listeners will be presented with auditory stimuli over speakers/headphones. The stimuli will be short utterances of emotional speech drawn from the LDC Emotional Prosody database (described in Section 2.1), which

Table 1: Subject Judgment Correlations

	ang	enc	hap	fru	int	anx	sad	con	bor	fri
angry										
encouraging	-0.30									
happy	-0.34	0.72								
frustrated	0.73	-0.37	-0.43							
interested	-0.04	0.63	0.58	-0.12						
anxious	0.18	-0.11	-0.12	0.28	0.11					
sad	-0.06	-0.31	-0.36	0.15	-0.32	0.16				
confident	0.06	0.52	0.43	-0.05	0.52	-0.20	-0.33			
bored	-0.07	-0.30	-0.32	0.07	-0.45	-0.18	0.37	-0.21		
friendly	-0.39	0.76	0.78	-0.43	0.59	-0.13	-0.25	0.44	-0.29	

Table 2: Acoustic Correlates of Emotion Types

feat	angry	anx	bored	conf	encour	friend	frust	happy	inter	sad
meanF0	0.217	0.242	-0.62	0.393	0.438	0.416	0.135	0.488	0.716	-0.364
F0event	0.199	0.282	-0.654	0.377	0.484	0.441	0.115	0.534	0.731	-0.303
meanRMS	0.237	0.255	-0.389	0.331	0.178	0.148	0.21	0.265	0.431	-0.256
maxRMS	0.261	0.263	-0.434	0.276	0.18	0.159	0.229	0.284	0.45	-0.201
rate	0.109	-0.033	-0.243	0.165	0.251	0.207	-0.015	0.201	0.266	-0.299

were used in our pilot experiment. Listeners will be instructed to simply view the set of faces while listening, and will not be explicitly asked to associate any particular face with the speech. Previous research using eye-tracking has shown that listeners tend to fixate objects which are mentioned or in some way related to what they are hearing, even if not explicitly told to do so (See, e.g., (Cooper, 1974)). Twenty native speakers of Standard American English will participate as listeners in the experiment.

Listeners' eye movements will be recorded and time-aligned with the speech input. Each face in the visual scene will be assigned a unique pixel region, and eye movements will be automatically coded with respect to these defined regions. This allows a log to be created indicating which face the listener was fixating at each point in time, at 5 millisecond intervals (the sampling rate of the eye-tracking device is 240 fps). The probability that a particular face will be fixated, given a particular emotional utterance, can thus be objectively determined. We will take this objective measure as an indicator of listeners' subjective decisions, in order to determine (i) whether listeners' earliest visual preference matches their final choice, (ii) how quickly listeners identify their final choice, (iii) which visual depictions are candidates before that deci-

sion and for how long, etc. These subjective and objective results will then be analyzed to identify salient prosodic and acoustic cues for identified emotional states.

Experiment 2: Incremental emotion recognition: This experiment is designed to elicit the point at which listeners can use various lexical and acoustic/prosodic information to reliably recognize a speaker's emotional state from an utterance.

The method for this experiment will be similar to Experiment 1, in that listeners will view a set of faces selected from the Ekman & Friesen (Ekman and Friesen, 1976) materials, while listening to auditory stimuli of emotional speech. The experiments differ in the nature of the auditory stimuli presented: Experiment 1 will use short dates and digit sequences from the LDC database, while Experiment 2 will use short sentences of emotional speech in which the location of a target lexical item evoking a certain emotion is systematically varied. In this way, we will be able to determine how both acoustic/prosodic and lexical information is integrated into emotional speech perception as an utterance unfolds. The following are two examples of utterance pairs which differ in the location at which the target lexical item (which might serve to convey anger here) appears.

This is just nonsense to me.
This is just a lot of nonsense to me.

I want to talk to a supervisor right now.
I want a supervisor right now.

In the experiment, both versions of each utterance pair will be uttered with the intention of conveying, for example, anger (other emotions will be used as well). Both the acoustic/prosodic information in the speech signal, as well as the target word, should lead listeners to fixate significantly more on the face which depicts anger than on other faces (as hypothesized by Experiment 1). The question specifically addressed in this experiment is the point at which fixation on the angry face become reliable, *in relation to the unfolding speech stream*. That is, are listeners able to reliably identify the emotion based on just the acoustic/prosodic information available from the start of the utterance, or must they wait until consistent lexical information is also encountered? We suspect that both factors will play a role in the perception of emotional state, though the eye fixations patterns will provide a quantitative measure of the trade-offs — and crucially the *time-course* of the tradeoffs — between these factors.

4 Emotion Detection from Perceptually Validated Cues

The data we obtain from these laboratory experiments will, we hope, enable us to build emotion recognizers from cues discovered in reliable experimental data. After determining empirically the combinations of features which reliably convey particular emotions to human subjects, we will build models making use of these cues to test on hand-labeled corpus data. If these tests are successful, we will be in a position to address the current dearth of corpora labeled by hand with emotion tags with automatic labeling. Even if our automatic labels must be post-processed, starting with a base hypothesis in this, as in the labeling of intonational features (Syrdal et al., 2001), should speed up the labeling process considerably.

References

Ang, Dhillon J., A. R., Krupski, E. Shriberg, and A. Stolcke. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2037–39, Denver.

Batliner, A., R. Huber, H. R. Niemann, E. Nöth, J. Spilker, and K. Fischer. 2000. The recognition of emotion. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast.

Bulut, M., S. S. Narayanan, and A. K. Syrdal. 2002. Expressive speech synthesis using a concatenative synthesizer. In *Proceedings of ICSLP 2002*, pages 1265–1268, Denver.

Burkhardt, F. and W. F. Sendlmeier. 2000. Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast.

Cahn, J. 1988. From sad to glad: Emotional computer voices. In *Proceedings of Speech Tech '88*, pages 35–36.

Cooper, Roger M. 1974. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6:84–107.

Cornelius, R. R. 2000. Theoretical approaches to emotion. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast.

Cowie, R. 2000. Describing the emotional states expressed in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, Belfast.

Cowie, R. and E. Douglas-Cowie. 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Proceedings of ICSLP-96*.

Ekman, Paul and W. V. Friesen. 1976. *Pictures of Facial Affect*. Consulting Psychologists Press, Palo Alto, CA.

Gussenhoven, C. 2002. Intonation and interpretation: Phonetics and phonology. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence.

Huber, R., A. Batliner, J. Buckow, E. Nth, V. Warnke, and H. Niemann. 2000. Recognition of emotion in a realistic dialog scenario. In *Proceedings of the International Conference on Spoken Language Processing*, pages 665–66, Beijing.

Kienast, M. and W. F. Sendlmeier. 2000. Acoustical analysis of spectral and temporal changes in emotional speech. In *ISCA Workshop on Speech and Emotion*, Belfast.

- Lee, C. M. and S. S. Narayanan. 2002. Combining acoustic and language information for emotion recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 873–76, Denver.
- Litman, D., J. Hirschberg, and M. Swerts. 2001. Predicting user reactions to system error. In *Proceedings of ACL-2001*, Toulouse.
- Mozziconacci, S. J. L. and D. J. Hermes. 1999. Role of intonation patterns in conveying emotion in speech. In *Proceedings of the XIV International Congress of Phonetic Sciences*, pages 2001–2004, San Francisco.
- Murray, I. R., M. D. Edgington, D. Campio, and J. Lynn. 2000. Rule-based emotion synthesis using concatenated speech. In *ISCA Workshop on Speech and Emotion*, Belfast.
- Pereira, C. 2000. Dimensions of emotional meaning in speech. In *ISCA Workshop on Speech and Emotion*, Belfast.
- Pollermann, B. Z. 2002. A place for prosody in a unified model of cognition and emotion. In *Proceedings of Speech Prosody 2002*, pages 17–22, Aix-en-Provence.
- Scherer, R. R. 2000. Psychological models of emotion. In *The Neuropsychology of Emotion*. Oxford University Press, pages 137–62.
- Schröder, M. 2001. Emotional speech synthesis: A review. In *Proceedings of Eurospeech 2001*, pages 561–564, Aalborg.
- Syrdal, A. K., J. Hirschberg, M. Beckman, and J. McGory. 2001. Automatic tobi prediction and alignment to speed manual labeling of prosody. *Speech Communication: Special Issue on Speech Annotation and Corpus Tools*, 33(1–2), January.
- Tanenhaus, Michael K., Michael Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.
- Yuan, J., L. Shen, and F. Chen. 2002. The acoustic realization of anger, fear, joy, and sadness in Chinese. In *Proceedings of the ICSLP 2002*, pages 2025–2028, Denver.