

Perceptions of Emotions in Expressive Storytelling

Cecilia Ovesdotter Alm, Richard Sproat

Department of Linguistics
University of Illinois at Urbana-Champaign, IL, U.S.A.

ebbaalm@uiuc.edu, rws@uiuc.edu

Abstract

Whereas experimental studies on emotional speech often control for neutral semantics, speech in naturalistic speech corpora is characterized by contextual cues and non-neutral semantic content. Moreover, the target emotion of an utterance is generally unknown and must be inferred by the listener. Within the context of having child-directed expressive text-to-speech synthesis as goal, we describe a perceptual study based on an expressive spoken corpus of children’s stories with unknown emotional targets, and report on interannotator agreement in a forced-choice discrimination task. Moreover, a threshold of high agreement was used to establish subsets of confident exemplar utterances for emotional classes, comprising 35% of the initial corpus. The exemplars were clustered based on the differences from the default mean *neutral* for 11 global acoustic features, yielding clusters cutting across emotion boundaries, some of which reflected arousal levels, with the *neutral* exemplars showing particularly complex distributions. Moreover, the mean features for four emotional exemplar categories were contrasted against the default, finding both expected and contradictory tendencies, compared to previous reports. The results indicate that semantic and prosodic cues collaborate to express and reinforce emotional contents, while emotional sequencing seems likely to be another factor which contributes to emotional perception in this domain.

1. Introduction

Emotions can be treated as categorical concepts, or can alternatively be described by dimensional or appraisal models; but despite the strong interest in emotion, there is currently no accepted definition for emotion, e.g. [5]. This study adopts the “big six” [5], i.e. *angry* (A), *disgusted* (D), *fearful* (F), *happy* (H), *sad* (Sa), and *surprised* (Su), adding *neutral* (N) as the default. These emotions have been investigated and characterized in terms of acoustic parameter settings for natural and synthesized speech and copy synthesis [14], [15], [3], [13].

However, studies on emotional speech often depend on portrayals of known target emotions that use semantically neutral sentences which lack context [15], whereas people also relate to non-vocal affect cues [16]. In naturalistic speech corpora, the target emotion of an utterance is generally not known and must be inferred by the listener, and the speech is also not devoid of contextual content or emotional meaning. In fact, [4] showed that spontaneous speech stimuli presented “in-isolation” versus “in-context” radically changed the emotional perceptions among listeners, whereas [11] demonstrated that lexical features improved automatic methods in binary emotion-detection, and [2] noted that more realistically-obtained emotional speech data increased complexity substantially.

We explore expressive prosody with the goal of child-directed text-to-speech synthesis for children with communication disorders [16]. The storytelling genre is rich in emotional content, and thus forms a good basis for exploring emotion in language and expressive prosody directed at children. In this study, we report on results from a perceptual experiment based on an existing corpus of children’s stories, interpreted freely by an experienced speaker in an expressive mode, thus lacking specific utterance targets. Section 2 describes the discrimination task performed by listeners at the utterance level. Section 3 reports on interannotator agreements computed with the *kappa* statistics in 3.1, whereas 3.2 explains how a threshold of high perceptual overlap in label-assignment was used to obtain a subset of confident exemplar utterances for different emotions. These were assessed in terms of their correspondence with previously reported acoustic emotion profiles, based on plotting category averages for 11 global acoustic parameters, as well as through clustering, reflecting the divergence from the *neutral* exemplar average. In section 4, we discuss how our results reflect the importance of non-acoustic cues for listeners’ perception of emotions, and we conclude noting the importance of emotional sequencing in 5.

2. Perceptual study

The perceptual study was conducted using the corpus described in [10], which consisted of two stories targeting young children, *Peter Rabbit* (PR) and *The Two Bad Mice* (2BM), read by a female speaker in an extremely expressive mode and then divided into utterance-length chunks.

The recordings were accessed online via a web survey which allowed for one-time submission. To avoid user-fatigue, each story was split into two surveys, representing its first and second half. After reading an informative text about the study, the participants completed the forced-choice task by listening to utterances in chronological story sequence, and marking one or optionally two labels that they felt characterized each utterance.¹ The participants also provided demographic information and reflected on their decision process.

The participants were UIUC students and included native English speakers (NS), a few early bilingual (B), and non-native speakers. Not all volunteers completed both parts. Since participants’ identities differed across surveys, each part was evaluated separately based on gender and native language (see table 1). Whereas the *all* group contained all participants, *NSB* con-

¹The option to mark two emotions was included because the interest lay in agreement trends, and distinguishing the more dominant emotion for an utterance was not always straight-forward. For example, 19% of the labeling events included two labels for the first half of PR.

Table 1: Number of participants per survey

Stimuli	All	M _{all}	F _{all}	NSB	M _{NSB}	F _{NSB}
PR-1	19	11	8	15	8	7
PR-2	14	5	9	10	2	8
2BM-1	27	12	15	12	4	8
2BM-2	26	14	12	12	6	6
Average	21.5	10.5	11	12.3	5	7.3

Table 2: Kappa statistics for normalized surveys

Stimuli	All	M _{all}	F _{all}	NSB	M _{NSB}	F _{NSB}
PR-1	.234	.242	.219	.242	.259	.243
PR-2	.241	.346	.19	.207	.246	.181
2BM-1	.375	.466	.334	.394	.518	.394
2BM-2	.496	.54	.444	.528	.557	.501

tained only the native and bilingual speakers.

3. Results

This section reports on the results of the survey in terms of interannotator agreement, and exemplar discovery and evaluation.

3.1. Interannotator agreement

Flammia’s tool [8] was used to compute the kappa statistics, after normalizing the survey data to one emotion label per annotator (by favoring the more frequent of the two marked labels for that particular utterance when an annotator had marked two emotions). Table 2 shows that the group with fewest annotators, the *male NSB* group, had the highest kappa for 3 out of 4 surveys, with the *male all* group having the highest score for the other survey.² We interpret the kappa score as an indicator of the difficulty of this perception task for a freely interpreted story corpus. Moreover, the range of the highest kappa scores can be seen as guidelines for expected accuracy bounds when applying automatic methods for emotional speech discrimination in the fairy tale domain.

3.2. Analysis of confident exemplars

As noted above, the corpus did not contain specified emotional targets. However, one could argue that high labeling agreement for a given utterance based on listeners’ perception signals an emotional target. Thus, utterances marked by extremely high agreement were defined as good emotional *exemplars*, and their acoustic properties were compared to the profiles suggested in the literature for each emotion class, to see if the high agreement on the exemplars depended on supported acoustic cues.

An exemplar was defined as an utterance for which both *all* and the *best kappa* group had $\geq 70\%$ agreement for a single emotion label. 46 utterances or 35% of the original corpus consisted of confident exemplars. Considering the limited size of the corpus, it is interesting that more than one third of the utterances were subject to such high agreement. As shown in

²Higher kappa was not necessarily obtained just by labeling more utterances as *neutral*; only in 1 of 4 cases did the group with the highest kappa scores tag more *neutral* labels. However, 2BM had more neutral label assignments overall.

Table 3: Intersecting exemplars for *best kappa + all* groups

A	D	F	H	N	Sa	Su
6	0	1	6	24	2	7

Table 4: Exemplars’ distribution across emotions and clusters

Label	A	F	H	Sa	Su	N
Cl-0						2
Cl-1			1	1	1	4
Cl-2				1		5
Cl-3	5	1	3		5	
Cl-4			1			8
Cl-5						4
Cl-6	1		1		1	1

table 3, the largest group were *neutral* sentences, but all other categories except *D* were also represented. However, [14] noted that *disgust* recognition depends more on facial expression.

F0, intensity and speech rate are commonly investigated acoustic correlates for emotions, e.g. [15], [14]. While *anger*, *fear*, and often *happiness* and *surprise* are generally characterized by varying degrees of increased speech rate, F0 and intensity values, *sadness* is marked by opposite behavior, and [6] found pausing a contributing cue to distinguish *fear* from *anger*. Moreover, [14] showed that some predictions for varying acoustic trends depending on the emotional intensity of the targeted emotion had been confirmed by empirical evidence while others had been contradicted. Thus, the following 11 global acoustic features were computed for each exemplar utterance using a python script:³

- F0: mean (*f0mn*), range (*f0range*), and standard deviation (*f0std*)
- Intensity: mean (*rmsmn*), range (*rmsrange*), and standard deviation (*rmsstd*)
- Speech rate: words/min (*wpm*), syllables/min (*spm*), and feet/min (*fpm*).
- Fluency: pause count (*pcnt*), pauses/min (*paupm*)

Two procedures were implemented to compare the confident emotion subsets to standard acoustic profiles reported in the literature. Both procedures were based on the assumption that the mean of the *neutral* exemplars represented a comparative default from which emotional expression would deviate either by increase or decrease in parameter values.

We used the `vcluster` tool [9] to cluster individual exemplars based on their percent difference from the default for each of the 11 acoustic features. If acoustic parameters determined perceptual label assignments, then members of a given cluster should correspond to a particular emotion label. The number of clusters reflected our initial set of emotion labels. Table 4 shows the resulting clusters and reveals that *neutral* exemplars were distributed across all clusters except cluster 3, and that there were no 1-to-1 correspondences between non-neutral labels having > 1 exemplars and particular cluster membership. A tentative attempt was made to describe and subjectively interpret the clusters, leaving out cluster 0 and 5 which were not

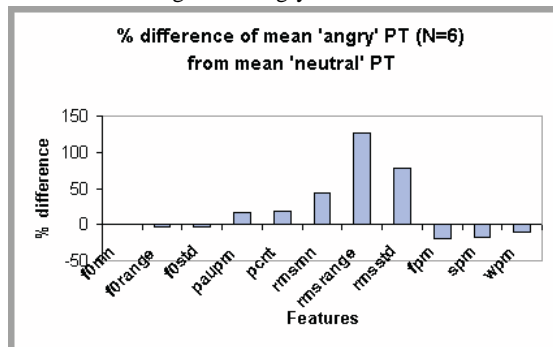
³The database from [10] had F0 values and transcriptions used to compute speech rate, whereas intensity values were obtained with *xwaves*. Zeros were ignored when computing F0 statistics.

intuitively interpretable, and ignoring the feature *fpm* which appeared unreliable.

- Cluster 1 was characterized by negative values, and reflected calm, slow, low arousal speech.
- Cluster 2 hinted at a slight decrease in *f0* and decreased pausing, with an increase in intensity and in speech rate. Most exemplars in this cluster were neutral and the speech reflected descriptive excitement.
- Cluster 3 only contained non-neutral exemplars and corresponded to high arousal speech. It showed a slight increase in *f0*, an increase in intensity and pausing and a decrease in speech rate.
- Cluster 4 was marked by slightly increased *f0*, slightly decreased intensity and speech rate scores, and increased pausing. It comprised mostly neutral exemplars, including list intonation. Subjectively, the speech reflected a soft low-level neutral or positive excitement.
- Cluster 6 was also marked by high arousal but fewer negative exemplars. It showed increased *f0* and intensity scores, and decreased speech rate. It differed from cluster 3 by displaying decreased pausing.

Next, the default *neutral* mean was plotted against the mean of exemplars corresponding to a given emotion label, given percent difference, for the acoustic features, cf. figures 1 to 4.⁴ The resulting plots were compared to profiles and results reported in the literature. As shown in fig. 1, the mean *angry* scores re-

Figure 1: Angry vs. default



flected, as expected, a fairly large positive increase in intensity. However, *angry* had slightly decreased *F0* scores and decreased speech rate, which contradicts its standard acoustic description. Additionally, *angry* showed increased pausing. Fig. 2 show that the mean *happy* scores had a slight positive increase in intensity and *F0*, however, speech rate decreased. Pausing also increased for *happy*. The mean acoustic features for *sad* did, as expected, generally decrease, compared to the default mean *neutral*, see fig. 3. Finally, as seen in fig 4, *surprised* showed increased intensity, and slight increase in *F0*, a slight decrease in speech rate, and unclear pausing patterns.

4. Discussion

The evaluation of the mean emotional classes presented above is limited because of the modest corpus size. Nevertheless, results indicate both expected and contradictory trends among the

⁴*Fearful* and *disgusted* were excluded because they had only one versus no exemplars, respectively.

Figure 2: Happy vs. default

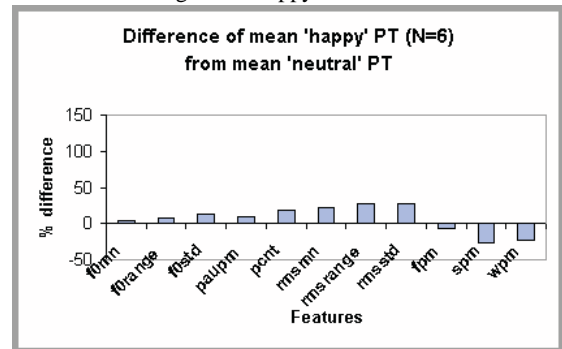
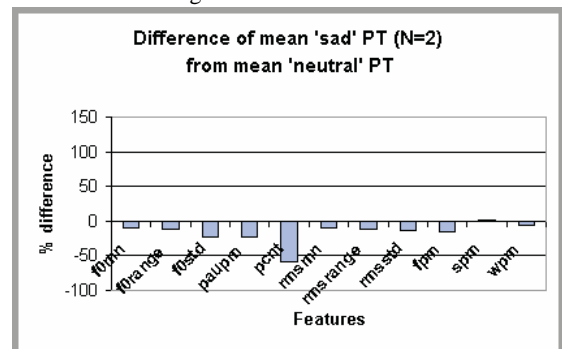


Figure 3: Sad vs. default



categories' acoustic features. In addition, clustering based on acoustic features revealed arousal levels. Indeed, *intensity* is the dimension of emotional similarity, compared to *valence* and *quality* [1], that is predominantly related to *F0*, intensity and speech rate [14]. Moreover, this dimension reflects distinctions within *emotion families*, such as *hot* versus *cold anger* [14].

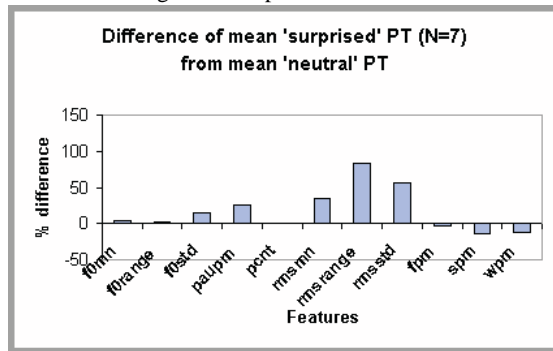
Given the successfully obtained set of exemplar utterances marked by high agreement, it appears that in addition to basic acoustic cues, non-acoustic features assisted listeners' perception of emotions. Closer inspection of the exemplars show that whereas neutral exemplars tended to be descriptive and lack lexis loaded with emotional meaning, lexical cues or broader semantic cues drawing on world knowledge indicated the emotional directionality for many non-neutral exemplars, e.g.:

1. *Angry*: Then Tom Thumb **lost his temper**.
2. *Fearful*: Peter sat down to rest; he was out of breath and **trembling with fright**, and **he had not the least idea which way to go**.
3. *Happy*: They squeaked with **joy!**
4. *Surprised*: **What a sight** met the eyes of Jane and Lucinda!
5. *Sad*: Peter began to **cry**.

In fact, several participants commented on the difficulty of keeping meaning and sound separated in the perceptual discrimination task, and as many as 93% (PR-2) vs. 81% (2BM-2) indicated that they considered utterance meaning as a factor when assigning emotion labels.

Separating meaning from prosody provides more controlled experimental settings, nevertheless it should be recognized that it is an artificial construct. Rather, semantic and prosodic cues

Figure 4: Surprised vs. default



tend to work in tandem and jointly convey emotional contents, so that non-neutral arousal levels in the acoustic signal support emotional directions present in meaning. This is in accord with the increased accuracy obtained after adding lexis by [11] in automatic binary emotion-detection. [12] noted differences in classification performance for combining acoustic and lexical features, but for human-human dialog data the combined approach mostly boosted performance. However, [7] found lower non-neutral identification rates in a perceptual tasks when subjects also listened to stimuli as opposed to when not, i.e. when they combined linguistic and prosodic information. But, the scope and target of their corpus was very different; consisting of actual agent-client interactions from a stock exchange call center, they hypothesize that politeness functioned as a control device in this context. Moreover, [7] reported encouraging results for detecting corpus-specific emotion labels automatically using lexical cues for an annotated corpus.

Finally, some studies ignored the *neutral* label in a forced-choice discrimination task, e.g. [3]. However, the results above showed that while *neutral* utterances were more often subject to high agreement, *neutral* exemplars clustered together with almost all non-neutral emotion classes. Moreover, the neutral exemplars represented a complex category with a wide variety of acoustic contours, including listing, contrastive emphasis, highlighting new versus given information, and so on. Thus, studies which ignore the neutral category could be intrinsically biased towards less confusion and undeserved high accuracy.

5. Conclusion

We presented empirical data from perceptual experiments on non-neutral contextualized expressive storytelling speech. Our target for emotional speech is child-directed text-to-speech synthesis. In addition to providing an analysis of expressive storytelling speech, and describing a procedure for obtaining subsets of confident emotion classes, the presented analysis highlighted the fact that perceptions of emotions in expressive speech depend also on semantic cues, in combination with acoustic parameter settings, and that high agreement for emotions was obtained even when some typical acoustic cues were absent.

As a further factor, additional preliminary results indicate that sequencing of emotions is highly relevant for fairy tales because they are highly structured. For example, fairy tales tend to consist of a neutral beginning, an emotional ascent, and a happy end. We are currently exploring how sequencing interacts with other features to determine perception of emotions.

6. Acknowledgements

We are thankful to Esther Klabbers and colleagues at Oregon Health and Science University for the corpus, and to Chilin Shih for xwaves training. This work was funded by NSF under award ITR-#0205731. The authors take sole responsibility for the work.

7. References

- [1] Banse, R. and Scherer K. R., "Acoustic profiles in vocal emotion expression", *Journal of Personality and Social Psychology*, 70(3):614–636, 1996.
- [2] Batliner, A. et al., "Desperately seeking emotions or: actors, wizards and human beings", *ISCA ITRW Speech and Emotion*, Newcastle, N. Ireland:195–200, 2000.
- [3] Cahn J., "The generation of affect in synthesized speech", *J. of American Voice Input/Output Society*, 8:1–19, 1990.
- [4] Cauldwell, R. T. "Where did the anger go? The role of context in interpreting emotion in speech", *ISCA ITRW Speech and Emotion*, Newcastle, N. Ireland:127–131, 2000.
- [5] Cowie, R. and R. R. Cornelius, "Describing the emotional states that are expressed in speech", *Speech Communication*, 40(1-2):5–32, 2003.
- [6] Devillers, L., Vasilescu, I. and L. Vidrascu, "F0 and pause features analysis for anger and fear detection in real-life spoken dialogs", *Speech Prosody Nara, Japan*:205–208 2004.
- [7] Devillers, L., Lamel, L. and Vasilescu, I. "Emotion detection in task-oriented spoken dialogs", *ICME*, 2003.
- [8] Flammia, Giovanni. Kappa coefficient tool. <http://www.theredesign.com/Technology/Dialogue/>
- [9] Karypis, G. CLUTO: A clustering toolkit. Technical Report: #02-017, 2003. <http://www-users.cs.umn.edu/~karypis/cluto/files/manual.pdf>
- [10] Klabbers, E. and van Santen, J. P. H. "Clustering of foot-based pitch contours in expressive speech". *5th ISCA speech synthesis workshop*, Pittsburgh:73–78, 2004.
- [11] Lee, C. M., Narayanan, S. S. and Pieraccini, R., "Combining acoustic and language information for emotion recognition", *ICSLP*:873–876, 2002.
- [12] Litman, D. J. and K. Forbes-Riley. "Predicting student emotions in computer-human tutoring dialogues", *ACL*: 351–358, 2004.
- [13] Murray, I. R. and Arnott, J. L., "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *J. Acoust. Soc. of America*, 93(2):1097–1108, 1993.
- [14] Scherer K. R., "Vocal communication of emotions: a review of research paradigms", *Speech Communication*, 40(1-2):227–256, 2003.
- [15] Schröder, M., "Emotional speech synthesis: a review", *Eurospeech*, Aalborg:561–564, 2001.
- [16] van Santen, J. et al. "Applications of computer generated expressive speech for communication disorders", *Eurospeech*, Geneva:1657–1660, 2003.