MANIPULATIONS OF F0 CONTOURS IN AFFECTIVE SPEECH ANALYSIS

Magdalena Oleśkowicz-Popiel, Jolanta Bachan

Institute of Linguistics, Adam Mickiewicz University in Poznań
magda.jastrzebska@gmail.com, jolabachan@gmail.com

Abstract: The study attempts to investigate the role of intonation contour in vocal expressions of emotions. For the needs of the study a dataset of utterances with manipulated intonation contour was created (F0 contour replacement). Manipulations were carried out using annotated and segmented speech recordings, including emotion portrayals, by professional actors. Segmentation of the recordings to syllables and phonemes was performed first automatically using a plugin of Annotation Pro software and then corrected manually. Original utterances for the manipulations were selected from a database of preannotated spontaneous dialogs between Police officers and citizens of Poznań (and surroundings) reporting an incident via 997 emergency call center. 3 utterances from each of 24 speakers, 12 males and 12 females, represented 6 emotional states (anger, irritation, fear, anxiety, sadness, shame). Next, 8 actors, 4 males and 4 females, were asked to produce the sentences in three ways: 1) neutral voice, 2) smiling voice, 3) exact imitation of the original sentence, supported by description of the situation and emotional profile of the speaker. A perception test carried out on a sample of actors' recordings and their F0 contour manipulations showed significant change in perception of emotions as a consequence of changes in intonation contour.

1 Introduction

The process of emotion communication via voice and speech consists of several important building blocks: speaker, listener, function of encoding and attribution (decoding) as well as distal and proximal cues. Affected by an emotion, speaker's voice production mechanism changes in complex ways to produce speech signal in which a given emotion will be encoded. This signal is acoustically a sum of such parameters as fundamental frequency, intensity, spectrum etc. Some of the changes convey a semantic message, others convey prosodic information. The latter one is understood as "distal cues". Next, the acoustic cues are modified by the listener's auditory system again with high complexity, and the listener perceives the emotion. The cues perceived by the listener are called "proximal cues" (e.g. pitch, loudness, voice quality) [1][2]. Numerous ways of approaching the research on affective speech, clear from such a model, is merely one of many challenges.

Type of data used for the research poses another problem, which is a choice between naturalness of the emotions and control over the collected data and experimental setting. The more control over the generated data the less spontaneity and naturalness of the expressed emotion and vice versa [3]. There are three main approaches for collecting emotional speech samples: 1) using actors for emotion portrayals, 2) inducing emotions via various scenarios and procedures, 3) recording spontaneous speech. Each has its own advantages and disadvantages, and the choice of the method is not always possible as it may be imposed by the aim of the research (e.g. emotion detection in call centers will mostly be based on spontaneous speech samples). In most cases researchers are able to choose a method, but the goal of the study may also determine the type of material dealt with. An exemplar overview of the existing database such as the one in [4] show explicitly predominant number of acted and elicited emotion databases over spontaneous speech databases.

Acoustic parameters that can be identified in the process of encoding emotions belong to either voice source parameters or articulation parameters. The first class encloses prosodic features and general voice quality (vocal effort and type of phonation) such as fundamental frequency measurements (mean, median, standard deviation, absolute minimum or maximum, absolute range, jitter, slope of F0 movement, F0 contour), intensity (mean, median, standard deviation, maximum, minimum, range, shimmer), speech rate and fluency (number of syllables per second, syllable duration, duration of accented vowels, number and duration of pauses, relative duration of voiced and unvoiced segments). The second class of parameters (refers to the filter part of the source-filter speech production model) on one hand is mandatory for the production of intelligible speech and on the other hand reflects extra-linguistic factors (such as facial expressions). The parameters in question are frequencies and bandwidths of formants as well as general distribution of energy at different frequencies in the spectrogram. The above listed acoustic cues can be assessed perceptually using terms such as: pitch (low/high), intensity (weak/strong), intonation (monotonous/modulated), instability (stead/shaky), roughness (not rough/rough), sharpness (not sharp/sharp), velocity of speech (slow/fast), articulation (bad/good), etc. [5]. Typically, a forceful and aggregative approach to finding acoustic correlates of emotions has been applied by many researchers in which as many as possible attributes are measured [6]. Not all the features are necessarily very informative or equally useful for emotion recognition and this approach does not explain the process of emotion inference from voice.

Predictions for effects of emotion on selected acoustic parameters have been made for example in [1] but except for a few research, the predictions have not been examined systematically. One of the features that have been studied but inconclusively is the intonation contour. In the literature there is no agreement as to the role and relevance of intonation patterns in the perception of emotions in speech [7][8].

In the present study a combination of three types of data have been proposed, namely spontaneous speech, actor portrayals and synthetic speech in order to validate actor portrayals and their re-synthesised versions as a valuable material for research on emotion perception and to investigate how intonation contour influences the perception of emotions.

2 Materials and methods

Materials used in the study were obtained in a stepwise process. First “997 corpus” of spontaneous speech dialogues has been constructed, then “actor portrayals corpus” has been recorded and annotated, and eventually F0 manipulations have been carried out on the “actor portrayals corpus”.

2.1 997 corpus

Database consisting of spontaneous real life dialogs via 997 emergency call center number between Police officers and citizens of Poznan (and surroundings) reporting an incident or asking for Police intervention was obtained by transcribing, segmenting and annotating the recordings. The reports were recorded in CCIT A-law format with 8 kHz sampling rate, with unknown characteristics of each of the telephone microphones (most likely different for each of the speakers). Therefore, although containing wide variety of natural, mostly negative affective states, the recordings are rich in background noises, differ in terms of their frequency spectra, and very problematic to obtain clear segmentation and F0 contour. For better control over experimental conditions “actor portrayals corpus” has been created based on selected recordings from “997 corpus”.

2.2 Actors‘ imitations

The utterances used as models for actor imitations were extracted from 24 recordings, 12 males and 12 females, for each sex 2 recordings evaluated as one of 6 negative emotions (fear, anxiety, anger, irritation, sadness, shame). Each utterance consisted of one or two prosodic phrases, with their semantic and syntactic content as close as possible to emotionally neutral.

Then 8 actors were asked to produce the chosen sentences in three different ways: 1) with natural, unemotional voice, 2) with cheerful, smiley voice 3) repeating phrases of each of the same sex speaker in a fashion as close to the original as possible (prosodically, emotionally and lexically). Before listening to phrases and repeating them one by one, actors familiarized with the broader context of the recordings and their emotional load. Actors were also given the written script of each of the phrases that they were to imitate but were instructed not to read them while speaking.

Recordings took place in an anechoic chamber of a professional recording studio, which ensured high audio quality and minimum background noise necessary to obtain reliable further measurements.

All of the utterances were transcribed phonetically and then segmented first automatically on to layers: “pho” – phonemes, “syl” – syllables using a plugin of Annotation Pro software [9] and then corrected manually. Next, segmentation was exported to a text file format using Annotation Pro export function.

2.3 F0 manipulations

For F0 contour replacement, a set of Praat and Python scripts was created. The first step was the F0 extraction from natural, emotional and smiling recordings using a Praat script. The Praat script extracts information from TextGrid files about phone labels on “pho” tier and the phone durations. Each phone duration is divided into three intervals from 0%-20%, 20%-80% and 80%-100% of phone duration and the mean pitch value is extracted for each of the intervals from a corresponding WAV file. The data from this step are saved in text files with “.F0” extension for each of the file in a directory (Table 1).

Table 1 – Examplar data from the F0 file, female voice, emotional recording.

Phone	Duration	Start time	Time 20%	Time 80%	End time	Mean F0 0%-20%	Mean F0 20%-80%	Mean F0 80%-100%
s	0.0892	0.3431	0.3609	0.4145	0.4323	undefined	undefined	undefined
k	0.0554	0.4323	0.4434	0.4767	0.4878	undefined	429.36	451.66
o	0.0645	0.4878	0.5007	0.5394	0.5523	460.014	450.03	489.57
Z	0.0732	0.5523	0.5670	0.6109	0.6256	436.69	373.13	429.97
y	0.0367	0.6256	0.6329	0.6550	0.6623	466.49	493.85	480.03
s	0.0700	0.6623	0.6763	0.7183	0.7323	459.49	480.92	undefined
t	0.0519	0.7323	0.7427	0.7739	0.7842	undefined	418.60	416.97
a	0.0580	0.7842	0.7959	0.8307	0.8423	381.68	368.52	407.55
w	0.0413	0.8423	0.8506	0.8754	0.8836	473.97	441.40	425.48

The next step is the creation of PitchTier files according to Praat format. Depending on what F0 contour replacement is to be done, a Python script extracts information about the phone duration from a “neutral” or “smiling” file and combines it with F0 values from a corresponding “emotional” file. The “emotional” F0 values are inserted in 10%, 50% and 90% of phone duration in “neutral” or “smiling” file/template. An excerpt of a PitchTier file is pre-

sented in Figure 1. The duration is taken from a “neutral” file and the F0 values are extracted from the “emotional” recording (cf. Table 1).

```
File type = "ooTextFile"
Object class = "PitchTier"

xmin = 0
xmax = 2.235929
points: size = 62
points [1]:
   number = 0.4520983
   value = 429.36
points [2]:
   number = 0.47262966
   value = 451.66
points [3]:
   number = 0.48331423
   value = 460.01
points [4]:
   number = 0.50552115
   value = 450.03
points [5]:
   number = 0.52772807
   value = 489.57
```

Figure 1 – PitchTier Praat file format, with combined data: “neutral” durations in seconds and “emotional” F0 values in Hz.

The final step is to replace the neutral pitch tier with the newly created pitch tier with the emotional F0 values and re-synthesise the neutral recording using the overlap-add synthesis in another Praat script. The different data used in the F0 manipulation process is presented in Figure 2. On top “neutral” recoding, in the middle “emotional” recording and at bottom the re-synthesised “neutral” recoding using overlap-add synthesis with the imposed “emotional” contour and anotation for the “neutral” recording.

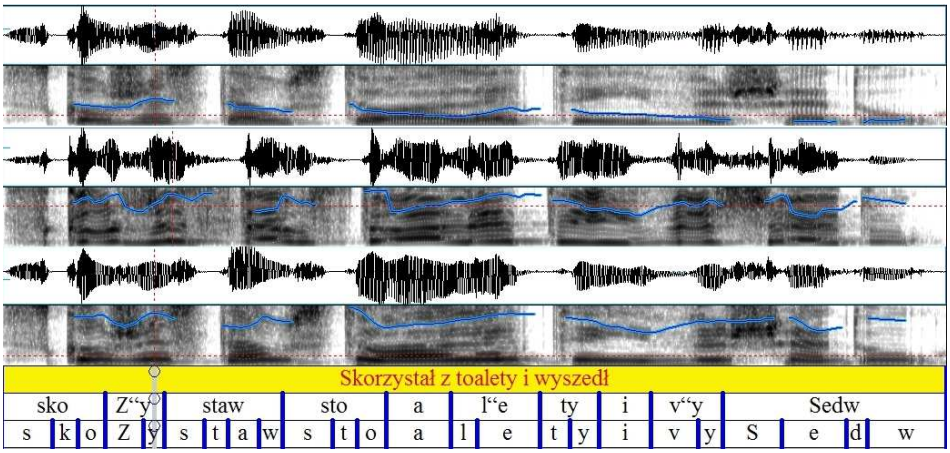

Figure 2 – F0 manipulation data - from the top: “neutral”, “emotional” and re-synthesised “neutral” recording with imposed “emotional” F0 values.

2.4 Perception test

For the perception test 6 sentences from the above described database has been chosen, each for different emotional state (Table 2). The sentences consist of one or two prosodic phrases and are emotionally neutral in terms of the semantic context.

Table 2 – List of sentences used in the perception test.

1. Skorzystał z toalety i wyszedł/ He used the bathroom and left. (FEAR)
2. To na pewno wiadomo / It is known for certain. (ANXIETY)
3. Na chodniku stają samochody/ Cars stop on the pavement. (ANGER)
4. Dostaliśmy się do lokalu/ We got into the room. (IRRITATION)
5. Dzwoniłam już do męża / I have already called my husband. (SADNESS)
6. Po prostu ta osoba już opuściła ten dom / simply this person has already left this house. (SHAME)

The whole test consisted of 24 utterances that is for each of the 6 sentences there have been 4 audio versions obtained: version 1 – natural voice in neutral state, version 2 – natural voice, the imitation of the original emotional state, version 3 – neutral natural utterance re-synthesised with emotional intonation contour, version 4 – smiling voice natural utterance re-synthesised with emotional intonation contour. Natural voice utterances came from one of the female actresses.

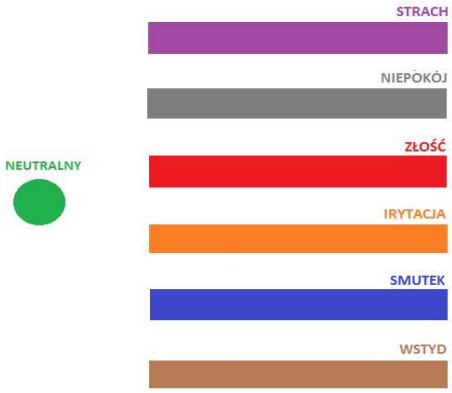

Figure 3 – Feature space used in the perception test; circle represents neutral state while stripes represent following emotions (from top): fear, anxiety, anger, irritation, sadness, shame.

The utterances have been presented via headphones to each of the 16 participants individually (males and females aged 20-24) in random order using *Annotation Pro* software perception test interface [9] . The participants were asked to listen to the recordings one by one and using a graphical representation of the feature space (Figure 3) to evaluate each utterance in terms of the emotion label and intensity (growing from left to right).

3 Results and discussion

Figure 4 and Table 3 show partial results of the perception test. They represent the perceptual evaluation of the actor portrayals of real emotions. In particular, emotions characterized by high activation such as fear, anger and irritation, have been properly identified in most cases. Anxiety, as a less intensive emotion, has been confused with other high-activation emotions.

Sadness has been confused systematically with fear which might result from both emotions being characterized by elevated mean F0 and narrow amplitude. Whereas shame has been systematically evaluated as sadness, which might be explained by the fact that recognition of shame is highly dependent on social context - perhaps without the context it is difficult to assert culturally grounded emotions.

Table 3 – Confusion matrix for version 1 utterances (actor imitations of emotions).

	fear	**anxiety**	**anger**	**irritation**	**sadness**	**shame**
fear	16	0	0	0	0	0
anxiety	1	8	3	4	0	0
anger	0	0	10	6	0	0
irritation	0	0	2	13	0	1
sadness	10	2	0	0	3	0
shame	0	0	0	0	14	1

Additionally, Figure 4 illustrates difficulties in percepting the intensity of emotions. In this test set two emotions with highest recognition rate that is fear and anger were also most agreed upon in terms of their intensity (high).

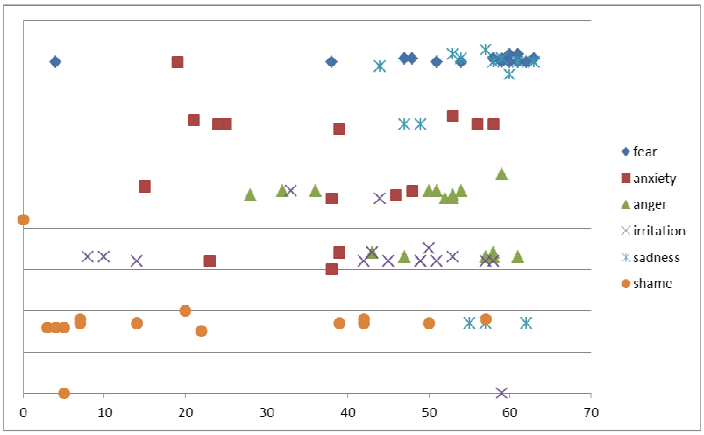

Figure 4 – Scatter plot of evaluations of version 1 utterances (actor imitations of emotions).

On one hand, these results may suggest disadvantages of using actor portrayals, but on the other hand they are a valuable indicator of individual differences in percepting emotions by individual listeners, and may shed a light on personal emotional sensitivity [10].

Since the focal point of the study has been to examine the perception not production of emotions, for the baseline of further verification of intonation contour influence on emotion perception not the emotional state intended by the actor, but the assessments of the listeners have been taken. As shown in Table 4, version 3 utterances (neutral natural voice utterance re-synthesised with emotional intonation contour from corresponding version 2 utterances) altogether have been evaluated 94 times (6 utterances x 14 listeners) out of which 58 evaluations were different from neutral while their version1 utterances were labeled neutral. After excluding from this summary cases when corresponding version 1 utterance were not labeled neutral, replacement of F0 contour resulted in changing evaluation from neutral to emotional in more than 90% cases. The effect of change over version 4 utterances (version 4 – smiling voice natural utterance re-synthesised with emotional intonation contour from corresponding version 2 utterance) was also strong – the change from smiley/neutral to emotional was

marked in 65 out of 94 cases which makes for almost 70%. This shows that in these cases intonation contour had predominant role over spectral characteristics of smiling voice.

Table 4 – Summary of change of Version 3 utterances evaluations (neutral natural voice utterance re-synthesised with emotional intonation contour from corresponding Version 2 utterance).

	Change to emotional	**No change**	**V1 not neutral**	
fear	21	0	6	
anxiety	6	2	2	
anger	11	1	3	
irritation	15	2	6	
sadness	4	3	10	
shame	1	0	1	
TOTAL	58	8	28	94

4 Conclusions

The short perception study showed that when searching for perceptual correlates of emotions it might be worthwhile having a closer look at intonation patterns, since straightforward F0 contour replacement had a significant role in changing perception from smiley/neutral voice to negative emotion label. It showed also that re-synthesis of natural voice gives a chance to systematically control and manipulate features subjected to validation in affective speech research.

5 Acknowledgements

The present study was supported by the Polish National Science Centre, project no.: 2013/09/N/HS2/02358, "*Vocal schemes of verbal emotion communication in linguistic perspective*".

6 References

[1] SCHERER, K.R.: *Vocal communication of emotion: A review of research paradigms*. Speech Communication, vol.40, pp. 227-256. 2003.

[2] SCHERER, K. R., & T. BÄNZIGE: *On the use of actor portrayals in research on emotional expression*. In K. R. SCHERER, T. BÄNZIGER, & E. B. ROESCH (Eds.) *Blueprint for affective computing: A sourcebook*. Oxford University Press, pp. 166-176. 2010.

[3] CALLEJAS, Z., R. LÓPEZ-CÓZAR: *Influence of contextual information in emotion annotation for spoken dialogue systems*. Speech Communication, vol. 50, pp. 416–433. 2008.

[4] VERVERIDIS D., C. KOTROPOULOS: *Emotional speech recognition: Resources, features, and methods*. Speech Communication, 48 (9), pp. 1162-1181. 2006.

[5] BÄNZIGER, T., K.R. SCHERER: *A study of perceived vocal features in emotional speech*. In: *Proceedings of Voqual 2003*, Voice Quality: Functions. Analysis and Synthesis. ISCA Workshop, pp. 169-217. 2003.

[6] SCHULLER, B., S. STEIDL & A. BATLINER: *The interspeech 2009 emotion challenge*. In Proc. *Interspeech*, pp. 312-315. 2009.

[7] MOZZICONACCI, S. J. & D. J. HERMES: *A study of intonation patterns in speech expressing emotion or attitude: production and perception*. IPO Annual Progress Report, 32, pp. 154-160. 1997.

[8] BÄNZIGER, T., & K. R. SCHERER: T*he role of intonation in emotional expressions*. Speech communication, 46(3), pp. 252-267. 2005.

[9] KLESSA, K.: *Annotation Pro [Software tool]*. Version 2.3.1.5. Retrieved from: http://annotationpro.org/ on 2016-02-17. 2016.
[10] BÄNZIGER, T., D. GRANDJEAN, & K.R. SCHERER: *Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT).* Emotion 9.5, pp. 691-704. 2009.