On the Role of Pitch in Perception of Emotional Speech

Noam Amir[1], *Eitan Globerson*[2]

[1] Department of Communication Disorders, Tel Aviv University, Israel
[2] Academy of Music and Dance, Israel

noama@post.tau.ac.il, gleitan@zahav.net.il

Abstract

Two experiments investigated the role of intonation in perception of basic emotions. In the first experiment, pitch contours of stimuli from a corpus containing portrayals of anger, joy, fear and sadness were manipulated with respect to range, mean and smoothness. In the second experiment, pitch contours of identical words portraying different emotions were exchanged. In each experiment, the emotional category and intensity of the original and manipulated stimuli were evaluated by two separate groups of 20 participants. Results of the first experiment show mainly that pitch mean and range should vary congruently to portray activation correctly, and demonstrate the interaction in varying these two parameters. Results of the second experiment show that a pitch contour conveying high activation is not sufficient in conveying the appropriate emotion, if the other paralinguistic cues are not also in accordance. A pitch contour indicating low activation, on the other hand, is apparently a more powerful cue and thus less reliant on other cues.

Index Terms: emotion, speech, intonation, perception

1. Introduction

It has long been acknowledged that pitch has a central role in production and perception of emotional speech [1], with more recent studies indicating that pitch mean and range are strong indicators of Activation or Arousal [2]. Many studies have shown, however, that other prosodic parameters such as intensity, timing and voice quality, are also involved in this process [3, 4,5]. It is therefore interesting to try and isolate the contribution of each of these factors to perception of emotion. Several previous studies have explored this issue using resynthesis, manipulating various prosodic or spectral properties and examining the effect on recognition scores [5, 6,7]. This is of interest for producing emotional speech in text-to-speech systems, however such tools can also be used studying the acoustic cues themselves.

In this study we present two experiments which attempt to investigate the contribution of pitch to emotion perception in a controlled manner. However we chose to perform manipulations on recorded emotional speech rather than text-to-speech synthesis. Both experiments employ a corpus of emotional speech which has been analyzed extensively in several other studies [8]. It is composed of short sentences and single-word (nonsensical and meaningful) emotional utterances with neutral content, recorded from 4 speakers and conveying 4 emotions: Anger, Joy, Fear and Sadness.

In experiment 1, the pitch contours of the different utterances were manipulated to exaggerate or diminish the pitch movements, in order to determine how this influenced the detected emotion and its intensity. In the second experiment, pairs of pitch contours from identical words uttered in different emotional expressions were exchanged, in order to examine which cues would be dominant: the pitch contour or the remaining acoustic cues.

The next section describes in some detail the corpus employed here, followed by detailed descriptions of each experiment, and a conclusion.

2. Emotional Corpus

A full description of the emotional corpus, detailing how it was recorded and evaluated can be found in previous work by the present authors [7]. In brief: stimuli were recorded in a professional recording studio by four professional actors (two female, and two male). The stimuli included nonsense monosyllabic utterances, nonsense polysyllabic words, Hebrew words and Hebrew sentences. None of the words and sentences had any linguistic emotional content. The stimuli represented four basic emotions: anger, joy, fear and sadness. The stimuli were validated by a panel of 20 independent judges, receiving an overall average recognition rate of 79.0% (SD=15.9%).

3. First Experiment

3.1. Objectives

The objectives of this experiment were to determine the effect of several properties of the pitch contour on the perception of emotion: 1) *Small pitch movements*: normally produced pitch contours contain major pitch movements known to signify both pragmatic and affective information, such as a final rise indicating a question. However, there are many smaller pitch movements which may have a significant but less obvious role [9]. Removing the small pitch movements, while retaining the large ones, may enable a controlled investigation of their perceptual importance. 2) *Pitch range*: this parameter has long been considered an important property of the pitch contour, usually considered to be an indicator of Activation or Arousal [2]. By stretching or compressing the pitch contour, the pitch range can be shifted considerably, without damaging its overall shape. 3) *Pitch mean*: This is also considered an important indicator of emotion, though it is not clear whether it is a global feature, independent to some degree of the speaker's normal mean, or must be normalized to each speakers individual mean. Once again, shifting the pitch mean can help determine its perceptual value.

3.2. Methods

In order to keep the listening task manageable, 92 utterances were selected from the original emotional speech corpus. The selection was balanced to ensure nearly the same number of utterance per emotional category (23 or 24 per category) and per utterance type (21-24). All the selected stimuli were originally identified correctly at a rate of over 68%. Each of these utterances was then manipulated in four ways:

10.21437/SpeechProsody.2014-19

1. Smoothing: removing any small pitch movements. This was carried out using Praat software's "stylize" manipulation with a parameter of 2 semitones.
2. Increase of pitch range by a factor of 2.
3. Decrease of pitch range by a factor of 2
4. Shifting of the pitch mean: Globally, pitch means for anger and sadness were low and means for joy and fear were high, both for men and women. Their overall averages were 260Hz for women's utterances, 167Hz for men's utterances. Thus, all stimuli were shifted to these mean values, for women and men respectively.

All manipulations were performed in Praat software. The net result was a collection of 460 stimuli: the original 92, and four manipulations of each.

The experiment was implemented as a Matlab GUI. For each stimulus, participants had to choose one of five emotions: anger, sadness, joy, fear or neutral, and rate the emotional intensity on a scale of 1 to 3. Despite the fact that the corpus did not contain neutral stimuli, the neutral category was included to account for cases in which the emotional content was possibly lost due to the above manipulations.

The participants were a group of 20 women, aged 20 to 30 (M=26). All participants had no reported hearing problems, 12-16 years of education, and were native Hebrew speakers.

3.3. Results

In the interest of brevity, only the main results of this experiment are presented here.

3.3.1. Identification of emotions

Table 1 shows the average recognition score per stimuli, for each emotion and each type of manipulation, in percents. Figure 1 shows the same results graphically. Evidently, baseline recognition scores (i.e. for non-manipulated stimuli) were very similar, over 80%, for all the emotion types.

Observing each row, it becomes obvious that the different types of manipulation had different effects on recognition of the various emotions.

The main findings from Table 1 are as follows: **Anger** recognition was reduced significantly ($p<0.05$) by shifting the pitch (in this case to higher levels) and *reduction* of the pitch range. **Joy** recognition was reduced significantly by pitch shifting (to lower levels in this case) and *reduction* in pitch range. **Sadness** recognition was reduced significantly by *increasing* the pitch range. **Fear** recognition was lowered greatly by pitch shifting (to lower levels).

*Table 1. Average **recognition** scores (%) per emotion and manipulation type. Asterisks mark statistically significant differences.*

Emotion	Original	Stylized	Shifted	0.5 x range	2 x range
Anger	84	86	73*	74*	80
Joy	81	74*	69*	53*	84
Fear	81	77	51*	76	71
Sadness	84	86	84	87	72*

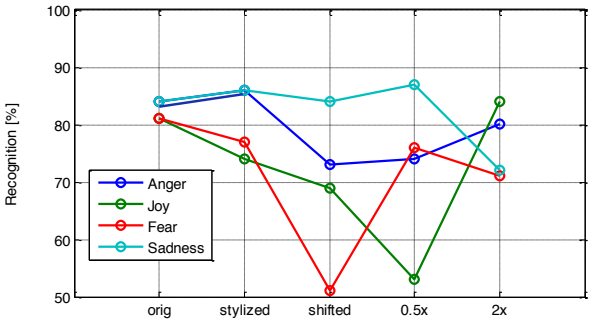

Figure 1: *Recognition scores of Table 1*

3.3.2. Ratings of emotional intensity

Table 2 shows mean intensity scores per emotion, for each manipulation. Anger and Joy scores fell when pitch range was reduced. Joy scores increased when pitch range was increased, whereas this had the opposite effect on Sadness. Finally, pitch shifting reduced intensity scores for Joy and Fear. This is in line with the fact that it reduced their recognition scores also.

*Table 2. Average emotional **intensity** scores per emotion and manipulation type (normalized to a 10-point scale). Asterisks mark statistically significant differences.*

Emotion	Original	Stylized	Shifted	0.5 x range	2 x range
Anger	6.1	6.3	6.1	5.4*	6.
Joy	5.7	5.6	5.2*	4.8*	6.6*
Fear	6.9	7	5.3*	6.9	6.5
Sadness	6.5	6.3	6.4	6.4	5.7*

3.4. Discussion

The results above show several interesting trends. In the original stimuli, the two emotions with high activation, Anger and Joy, demonstrated opposing tendencies with regard to average pitch. Thus, shifting them to the overall average (reducing pitch in Joy and increasing it in Anger) caused recognition rates to fall for both. As activation is commonly acknowledged to be associated with high pitch range [Banziger], recognition rates of both these emotions also fell when pitch range was reduced.

Sadness and Fear also employed opposing tendencies in the original corpus, with regard to average pitch: Fear being high and Sadness low. However, shifting the average pitch had a large effect on Fear, but no effect on sadness. Nevertheless, since these emotions have relatively low activation, *increasing* the pitch range for these emotions resulted in lower recognition rates.

Pitch stylization had some interesting effects also. It affected Joy most significantly, reducing recognition rates. This may indicate that recognition of Joy relies to some extent on small pitch movements as well as large ones. On the other hand, stylization increased recognition of Anger and Sadness slightly, though not significantly. Possibly this type of manipulation removed some small pitch movements that might in fact cause the emotion to sound more ambiguous in these cases.

The trends for intensity scores are roughly similar to those found for recognition scores. Emotions with high activation (Anger and Joy) were adversely affected by *reducing* pitch range, though less affected by *increase* in range. Pitch shifting had a similar effect on intensity as on recognition, though not in all emotions.

Overall, this experiment reveals that pitch mean and pitch range behave somewhat independently. Range appears to signify activation. Thus, reducing the range for emotions with high activation (Anger and Joy) appears to confound the listeners. However, reducing the pitch range for emotions with low activation (Sadness and Fear) does not have an identical effect, only marginally changing the recognition of both.

Increasing the range had no effect on emotions that had a large pitch range to start with, but reduced the recognition rates for emotions with low activation (Fear and Sadness) where a large pitch range was not expected.

Shifting the pitch mean adversely affected nearly all emotions, probably because shifting was done towards the overall average, thus reducing the degree of emotional content. This raises the question whether listeners were able to normalize their expectations of average pitch to the speakers, or whether speakers are preconditioned towards some overall global average in this respect. Interestingly, Sadness was not at all affected by this manipulation.

To summarize, the results show that the parameters of pitch range and mean and their interactions have important significance in production and perception of emotions, which are highlighted in this experiment.

4. Second Experiment

4.1. Objective

The objective of this experiment was to examine the degree to which the specific pitch contours associated with each emotional production were responsible for the correct perception of the emotion.

4.2. Methods

The original corpus was scanned for utterances which could be "paired": i.e. words or nonsense words uttered by the same speaker while expressing two or three different emotions. 104 such utterances were found. For these utterances, the pitch contour of one emotion could be synthesized onto the utterance the other one or two remaining utterances. The final corpus was thus composed of:

1. 104 original utterances
2. For pairs of utterances uttered in *two* different emotions, the pitch contour of one utterance was synthesized onto the other utterance, and vice versa, producing two new stimuli.
3. For triplets of emotions uttered in *three* different emotions, each combination of cross synthesizing pitch contour and utterance were created, producing six new stimuli

The net result was a corpus of 268 stimuli: 138 nonsense words (54 original and 84 synthesized), and 130 Hebrew words (50 original and 80 synthesized).

All manipulations were performed in Praat software. This experiment was also implemented as a Matlab GUI. For each stimulus, the participant had to choose one of five emotions: anger, sadness, joy, fear or neutral, and rate the emotional intensity on a scale of 1 to 3.

The participants were a group of 20 women, aged 20 to 30 (M=26.4), separate from the group that had participated in the first experiment. All participants had no reported hearing problems, 15-19 years of education, and were native Hebrew speakers.

4.3. Results

The average recognition score for all non-manipulated stimuli was 15.2 out of the possible 20, which is slightly lower than in experiment 1. The entire confusion matrix for these stimuli appears in Table 3. All values on the diagonal are in the vicinity of 15, with a relatively even distribution of errors. This indicates that on average, the emotions in the original, non-manipulated stimuli were fairly easy to recognize.

Table 3. *Confusion matrix of recognition scores for the original stimuli (%)*

Perceived / Produced	Anger	Joy	Fear	Sadness	Neutral
Anger	75	7	2	1	14
Joy	12	77	2	1	8
Fear	1	1	80	18	1
Sadness	7	0	10	72	11

For the manipulated stimuli, two confusion matrices can be calculated: one for identification of the stimuli according to the original emotion of each stimulus, and the second according to identification of the emotion represented by the synthesized pitch contour of each stimulus. These two matrices are presented in tables 4 and 5.

Table 4. *Confusion matrix of recognition scores for the manipulated stimuli, identified in accordance with the original emotions (%).*

Perceived / Produced	Anger	Joy	Fear	Sadness	Neutral
Anger	36	15	21	17	11
Joy	13	36	14	15	22
Fear	5	4	47	34	11
Sadness	7	5	28	46	13

Table 5. *Confusion matrix of recognition scores for the manipulated stimuli, identified in accordance with the synthesized pitch contour (%).*

Perceived / Produced	Anger	Joy	Fear	Sadness	Neutral
Anger	10	19	25	24	22
Joy	23	18	25	21	13
Fear	12	14	39	29	6
Sadness	18	11	20	36	16

Table 4 indicates that on average, changing the pitch contour caused recognition rates to fall, though they remained above the chance level (4/20) for all the emotions. Emotions Fear and Sadness remained more immune to changes of the pitch contour.

The first two values on the diagonal of Table 5 are very low. This indicates that imposing a pitch contour of Anger or Joy on utterances that originally conveyed other emotions, had very little success in causing these utterances to "shift" towards Anger or Joy. On the other hand, the second two values on the diagonal of this table are much higher, well above chance. This indicates that imposing the pitch contour of Fear or Sadness had a clear effect in shifting the perceived emotions towards these two.

Finally a more detailed analysis was carried out. Each combination of original and F0 emotion was analyzed separately, scoring perception according to 1) the original utterance; 2) the pitch contour. Results are presented in Table 6 and Figure 2.

Table 6. *Recognition scores for each cross-manipulation of original emotion and F0 contour emotion (%). Four blocks of rows are denoted with shading: hi activation to hi, hi to low, low to hi and low to low*

Row #	Original emotion	F0 emotion	Score according to original	Score according to F0	N
1	Anger	Joy	48	26	16
2	Joy	Anger	49	14	16
3	Anger	Fear	23	40	17
4	Anger	Sadness	38	28	11
5	Joy	Fear	35	32	11
6	Joy	Sadness	23	28	14
7	Fear	Anger	49	5	17
8	Fear	Joy	52	10	11
9	Sadness	Anger	50	12	11
10	Sadness	Joy	37	13	14
11	Fear	Sadness	40	52	13
12	Sadness	Fear	52	43	13

Some interesting observations can be made based on this table. For example, rows 1 and 2 show that emotions with high activation (Anger and Joy) are not easily "taken over" by swapping each other's pitch contours. However, rows 11 and 12 show that the opposite is true for emotions with low activation. Rows 2 - 6 indicate that Anger and Joy are more readily taken over by pitch contours of Fear and Sadness. Rows 7 – 10 show that Fear and Sadness are *not* easily taken over by Anger and Joy.

4.4. Discussion

The observations regarding the confusion matrices strengthen the notion that pitch is an important cue of emotion, though not the sole one. Imposing a "wrong" pitch contour reduced recognition score of the original emotion by approximately half, though emotions with high activation (Anger and Joy) appear to be more severely affected than emotions with low activation (Fear and Sadness).

Conversely, when looking at the same manipulated stimuli in order to observe when the imposed pitch contour "took over" the perceived emotion, a gross asymmetry can be observed. On average, the pitch contours of the highly active emotions had no discernible effect at all in taking over the perceived emotion. On the other hand, pitch contours of the emotions with low activation were moderately successful in taking over the perceived emotion.

Our conjecture is therefore that a pitch contour conveying high activation is far from being sufficient in conveying the appropriate emotion, if the other paralinguistic cues are not congruent. A pitch contour indicating low activation, on the other hand, is apparently a more powerful cue and thus less reliant on other cues.

The detailed analysis in Table 6 also raises some interesting asymmetries, showing that pitch contours can more easily draw one emotion to the other than the reverse. Line 3 shows that Anger is rather easily turned into fear, for example, but line 7 shows that the opposite is not true. Joy is easily converted to Sadness (line 6), but again the opposite is not true (line 10).

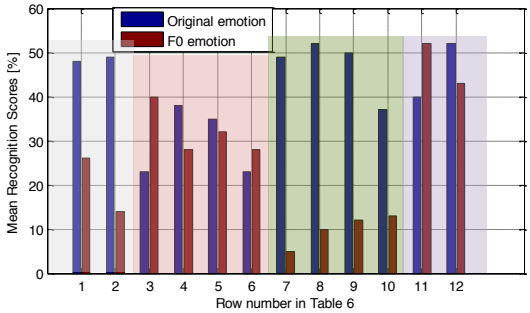

Figure 2: *Recognition scores of Table 6*. Groups of bars are color coded to fit the rows in the table.

5. Conclusions

Arriving at a clear and definite picture of the cues behind the production and perception of emotions is a daunting and likely impossible task. However, studies such as the present one can provide insight into the relative perceptual importance of these cues. Pitch has been long been acknowledged to be a central cue in emotion recognition, however this study quantifies its role to a certain extent, and underlines the fact that it has different influence on the perception of different emotions. The corpus used here offers a unique opportunity to study such effects in isolation, an opportunity which is not afforded by many emotional corpora. Presumably, manipulations of further cues, such as voice quality, could lead to results with even a higher degree of refinement in separating the effects of different cues to emotional perception.

6. Acknowledgements

The authors would like to thank Nina Gilad, Amit Lavi, Michal Rubinstein and Ravit Tahar for their assistance in running the experiments.

7. References

[1] Murray, I.R., Arnott, J.L., "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", Journal of the Acoustic Society of America. 93(2), 1097- 1108, 1993

[2] Banziger, T., Scherer, K.R., "The role of intonation in emotional expressions", Speech Communication, 46(3), 252-267, 2005

[3] Banse, R., Scherer, K.R., "Acoustic profiles in vocal emotion expression", Journal of Personality and Social Psychology, 70(3), 614-63, 1996.

[4] Yanushevskaya, I., Gobl, C., & Ní Chasaide, A., "Voice quality and loudness in affect perception", Proceedings of Speech Prosody 2008, Campinas, Brazil

[5] Gobl, C., Ni Chasaide, A., "The role of voice quality in communicating emotion, mood and attitude", Speech Communication 40(1), 189-212, 2003.

[6] Lindh, J., "A model based experiment towards an emotional synthesis", Proceedings of FONETIK 2005, Goteborg.

[7] Burkhardt, F., "Emofilt: the simulation of emotional speech by prosody-transformation", Proceedings of Interspeech 2005, Lisbon

[8] Globerson, E., Amir, N., Golan, O., Kishon-Rabin, L., Lavidor, M., "Psychoacoustic abilities as predictors of vocal emotion recognition", Attention, Perception and Psycophysics, 75(8), 1799-1810, 2013

[9] Lieberman, P., & Michaels, S. B., "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech", Journal of the Acoustical Society of America 34, 922–927, 1962