# Some non-F0 cues to emotional speech: An experiment with morphing

*Donna Erickson[1], Takaaki Shochi[2], Caroline Menezes[3], Hideki Kawahara[4] and Ken-Ichi Sakakibara[5]*

[1]Showa Music University, Kawasaki City, Japan, [2] Gipsa-Lab, Grenoble, France, [3]Goa, India, [4]Wakayama University, Wakayama, Japan, [5]Health Sciences University of Hokkaido, Japan
[1]EricksonDonna2000@gmail.com, [2]shochi38@gmail.com, [3]caroline_menezes2002@yahoo.com, [4]kawahara@sys.wakayama-u.ac.jp, [4]kis@hoku-iryo-u.ac.jp

## Abstract

This paper investigates some non-F0 cues to emotional speech. Two speech samples were collected from spontaneous speech: the word "leave"--one sample spoken with emotion (*sad*) and the other, as not-emotional. Using the morphing algorithm of STRAIGHT [1], we morphed a series of 12 utterances, starting from the non-emotional "leave" to the emotional "leave", keeping F0 at 300 Hz. Perception test results show that the morphed speech sounds could be identified as *sad*, with stimulus 12 being heard as most emotional. The results of a simple correlation, together with a PCA analysis of listeners' perceptual behavior, suggest that formant frequencies, specifically, lowering F2, F3, and F4 are important cues for perception of emotional (*sad*) speech.

## 1. Introduction

A number of factors are involved in perception of emotion/attitude of speaker, such as F0, intensity, duration, and voice quality, with F0 being an extremely strong cue to emotion. A recent study by [2] showed that utterances with higher F0 and larger amplitude are perceived as having high emotional intensity. In this paper, we wish to investigate the non-F0 cues, specifically, "voice quality/voice timber" cues of emotion/attitude. One definition of voice quality/timbre is "the quality of a sound by which a listener can tell that two sounds of the same loudness and pitch are dissimilar" [3]. Current investigation into voice quality has focused on measures of breathiness, such as H1-H2 or H1-A3, where H1, H2, and A3 are the amplitudes (dB) of the fundamental frequency, the second harmonic, and the harmonic peak associated with the 3rd formant, respectively. H1-H2, reflects the amount of glottal opening during the vibratory cycle; H1-A3, the speed of the closing phase of the vocal cycle e.g., [4], and for both, increasing positive values represent increasing breathiness. However, voice timbre/quality has many facets, as evidenced, for instance, by research on singing. The voice qualities of opera, belting, twang, and sob are outlined by [5], and involve a complex interaction of vocal fold mass and laryngeal settings, among other things. Specifically, for the sob (cry) quality, the larynx is said to be low and the aryepiglottic region, relaxed.

## 2. Methods

### 1.2. Data collection

Acoustic recordings (in conjunction with articulatory recordings using the 2D EMA system) were done at NTT Research Laboratories, Atsugi, Japan for an American (Midwest dialect) female speaker. This was conducted as an informal spontaneous telephone dialogue with another speaker (conversation partner) through an earphone/microphone set-up, where the other speaker sat in a separate room from the subject. The conversation partner asked the subject unrehearsed questions related to the subject's personal life to evoke spontaneous *happiness*, *sadness* or *anger*. The timing of the experiment was fortunate for collecting *sad* (grieving) emotions (including crying while speaking), since the subject was mourning the loss of her mother. A second set of control data utterances were also collected 5 months later, in which the speaker imitated her original utterances in a number of ways: (1) imitating the wording, phrasing, intonation and emotion of the original utterances while at the same time listening to a taped recording through headphones, and looking at a transcript; (2) imitating just the wording, phrasing and intonation (but not emotion), also while listening to the original utterance and looking at the transcript; and (3) reading the original utterance from the orthographic transcript of the original recording. A version of this paper appears in [6].

### 1.1. Stimulus morphing

The word "leave" was selected since this word was spoken with extreme *sad* emotion. We also selected the corresponding "leave" uttered in the control data utterances, spoken without emotion, but imitating the phrasing, and intonation pattern (but not average F0). Perception tests [6] showed that the "emotional leave" was given a rating of about 3.8 (on a scale of 1 to 5, with "5" indicating most "sad"), whereas the "leave" imitating phrasing and intonation pattern but not emotion was given a rating of about 1.2. We used STRAIGHT to morph a series of 12 utterances, starting from the imitated (non-emotional) "leave" to the emotional "leave". STRAIGHT decomposes the input speech signal into three parameters; fundamental frequency trajectory, aperiodicity spectrogram and smoothed spectrogram. The morphing procedure interpolates these parameters for the two sample of original speech. A time-frequency alignment was performed prior to this parameter interpolation based on manually assigned landmarks on each spectrogram. The default method of interpolation is linear interpolation of logarithmic parameters, in other words geometrical interpolation of the original linear parameters. The time-frequency alignment deforms the temporal and the frequency axes. This deformation effectively interpolates temporal cues (such as voice onset time) and frequency cues (such as formant frequency). Modifications on the frequency axis, the temporal axis and spectral level, and fundamental frequency can be controlled independently. In the current experiment, spectrogram and aperiodicity parameters were morphed in the same amount while the timing cues and the fundamental frequency trajectory were kept constant (3000 Hz) by taking

advantage of this independent morphing rate control capability.

### 1.2. Perception tests

There were 12 morphed stimuli, 3 randomizations, for a total of 36 utterances, plus a practice test of 5 utterances. The tests were administered through HDA200 Sennheiser headphones in a quiet room, using a Windows-based computer software from Runtime Revolution. The listeners, 78 Midwestern American college students from Ohio and South Dakota, responded to two questions: (1) rate each word according to the perceived degree of emotion on a 5 point scale, with "5" most *emotional*; (2) identify the perceived emotion—(1) *happy*, (2) *sad*, (3) no emotion, (4) other (5) unknown. The questions were framed to not bias the listeners' perception to a single particular emotion.

## 2. Results

### 2.1. Perception test results

Cronbach's Alpha showed that the homogeneity of distribution of responses for both questions are satisfied (0.98 for Q 1, 0.96 for Q 2). The results to "How emotional was the sound?" are shown in Figure 1. The repetitive effect was absent for the results, and we take the average of the 3 repetitions. There is an increase in rating of emotion from stimulus 1 to stimulus 12, with stimulus 12 being heard as most emotional. The graph suggests a "cross-over" from not-emotional to emotion at around stimuli 10, 11, and 12, which were all perceived at a level "3" emotion-- a value of "3" indicated the listeners heard the sound as "emotional." Such a morphing effect was justified by ANOVA ($F=22.60$, $p.<0.01$). Figure 2 shows the mean of recognition rate for the second task which asked the identification of appropriate emotion among 5 choices. While most listeners perceived correctly that the stimuli were *sad* there is also an increase of sad identification starting from stim8. This is concomitant with a decrease in decrease of identification rate for "no emotion" at the same point. This implies that some loss of recognition score for "no emotion" moved to *sad*. It is also important to note that accounting for only the good identification values (top line, with square boxes, indicating *sad*), the morphing effect was justified by ANOVA ($F=6.26$, $p.<0.01$). These distributions of mean values for both questions are highly correlated ($r = 0.81$, $p.<0.01$).

.

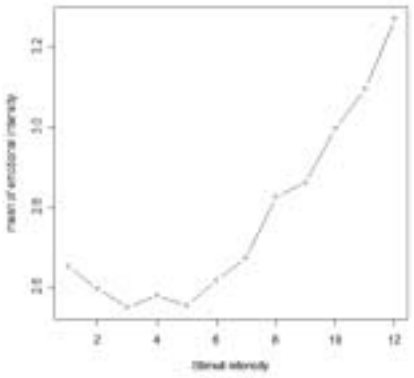

Figure 1*: Perception test results to question 1. "How emotional was the utterance." X-axis shows stimuli number with stimulus 1 representing morphed non-emotional "leave", and stimulus 12, emotional "leave". Y-axis shows mean intensity of perceived emotion by 78 American English listeners, 5 indicating "extremely emotional" and "1", "not emotional at all".*

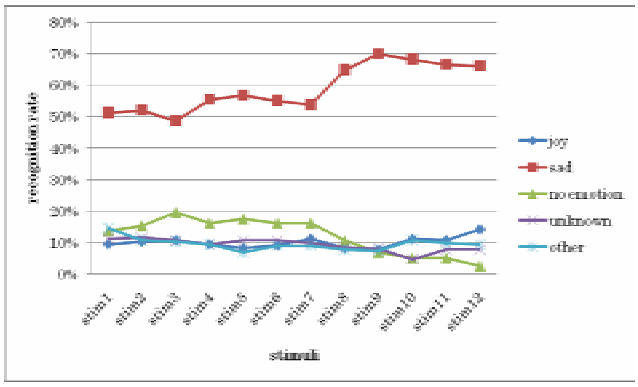


Figure 2*: Perception test results to question 2 "What emotion did you hear?". X-axis shows stimuli number with stimulus 1 representing morphed non-emotional "leave", and stimulus 12, emotional "leave". Y-axis shows recognition rate of perceived emotion.*

### 2.2. Acoustic analysis

Acoustic measurements of duration, F0, formants frequencies (F1, F2, F3, F4), and voice quality (H1-H2, H1-A3, spectral tilt) were made of the 12 morphed stimuli, as well as for the two original stimuli (the non-emotional "leave" and the emotional "leave"), using WaveSurfer and PRAAT, and are shown in Table 1.

F1 is highest for stimulus 12 (the morphed "leave" stimulus perceived best as emotional) and lowest for stimulus 1 (the morphed "leave" stimulus least-well perceived as "emotional") while the other formants (F2, F3, and F4) are lowest for stimulus 12 compared with stimulus 1. A similar pattern of lower F2, F3, and F4 is seen for the original emotional utterance compared to the original non-emotional one; however, F1 is actually lower for the original emotional utterance compared to the original non-emotional utterance. In terms of voice quality measurements, stimulus 12 (as well as the original emotional utterance), compared with stimulus 1 (and the original non-emotional utterance), have both lower H1-H2 and H1-A3 values, indicating that the emotional stimuli (both the original and the morphed one) were less breathy than the non-emotional stimuli both the original and the morphed one).

Fig. 3 shows the spectra of the original emotional utterance (dark line) compared with that of the original non-emotional utterance (light line); Fig, 4, the morphed stimulus #12 (dark line) compared with that of the morphed stimulus #1 (light line). For both the original *sad* speech and the morphed stimulus 12 (dark lines), we see F2, F3, and F4 low compared with original non-*sad* speech and morphed stimulus 1 (light lines). Also, we see a high peak of energy around 4-5 kHz in both the original and the morphed emotional utterances (black lines) which is missing in the original and morphed non-emotional speech. And, we see a trough of

energy around 5 khz, indicated by a circle in the figures, for both the original and morphed emotional utterances (dark lines), not seen for the non-emotional utterances (light lines).

Table 1: *Acoustic Measurements. "File name" indicates the morphed stimuli number, and the last two entries, the original non-emotional (NE) and emotional (E)utterances.*

| File Name | F1 | F2 | F3 | F4 | H1-H2 | H1-A3 |
|---|---|---|---|---|---|---|
| 1 | 297 | 2674 | 3276 | 4597 | 27.0 | 39.3 |
| 2 | 298 | 2671 | 3427 | 4497 | 26.0 | 37.3 |
| 3 | 299 | 2665 | 3352 | 4528 | 25.2 | 39.5 |
| 4 | 300 | 2657 | 3354 | 4480 | 24.2 | 39.8 |
| 5 | 301 | 2658 | 3340 | 4483 | 23.7 | 36.8 |
| 6 | 302 | 2640 | 3329 | 4478 | 22.7 | 36.9 |
| 7 | 304 | 2644 | 3326 | 4470 | 22.7 | 37.7 |
| 8 | 305 | 2515 | 3324 | 4480 | 23.1 | 36.0 |
| 9 | 307 | 2446 | 3333 | 4491 | 18.4 | 35.1 |
| 10 | 304 | 2396 | 3231 | 4409 | 20.4 | 38.2 |
| 11 | 307 | 2380 | 3179 | 4389 | 18.7 | 37.0 |
| 12 | 310 | 2364 | 3115 | 4382 | 20.0 | 36.9 |
| NE | 336 | 2764 | 3515 | 4617 | 49.2 | 86.6 |
| E | 318 | 2507 | 3161 | 4379 | 30.1 | 34.2 |

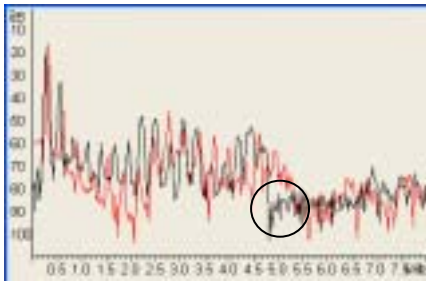

Figure 3*: Spectra of original emotional speech (dark) and non-emotional speech (light).*

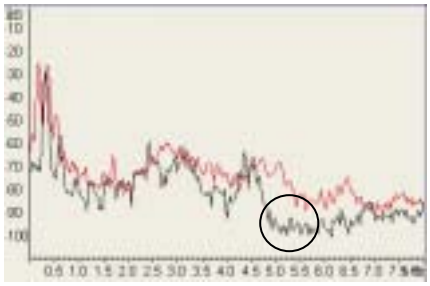

Figure 4*: Spectra of morphed stimulus 12 (dark) morphed stimulus 1 (light)*

### 2.4. Correlation with acoustics and perception

A Pearson correlation of intensity of perception of emotion with the acoustic measurements was done, and a post-hoc Bonferroni test of significance showed significant negative correlations with perception and F2, F3, and F1-F2. As F2 or F3 become lower, perception of emotion increases; also, as F2-F1 becomes smaller, perception of emotion increases. No significant correlation between perception and any of the voice quality measurements having to do with breathiness was found.

However, a simple correlation between acoustic parameters and perceived intensity of emotion may not be the best analysis approach, since the acoustic parameters are inter-related; hence, we also did a Principal Component Analysis (PCA), which can possibly reveal which component is more linked to listeners' perception behavior (see Fig 5 below).

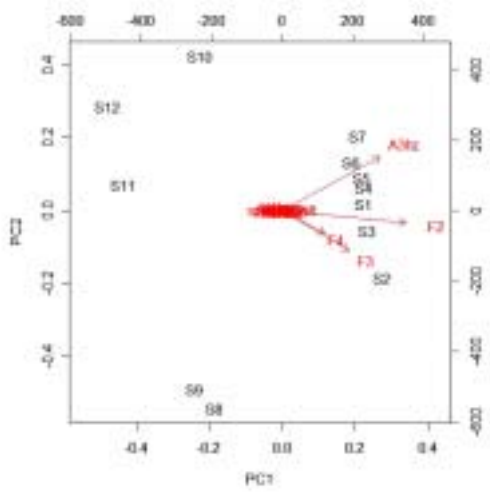


Figure 5*: Both listeners' perceptual behavior points for 12 step (from stim1 to stim12) and 15 acoustic parameters (duration, f0, f1, f2, f3, f4, H1hz, H1db, H2hz, H2db, A3hz, A3db, H1-H2, H1-A3 and spectral tilt) for 12 stimuli are presented on two dimensions according to Principal Component Analysis (PCA).*

According to Figure 5, two perceptual categories (category of stim1-stim7 on the right side of figure, and another of stim8-12 on the left side of figure) are observed. Concerning acoustic parameters, almost all acoustic parameters are concentrated on the center of the figure except A3hz, F2, F3 and F4 which are located on the right side of the figure. This means these 4 parameters show somewhat different variation of acoustic changing rather than the others. Moreover, this figure showed that the four acoustic parameters of A3hz, F2, F3 and F4 are related to the perceptual behavior of listeners to stim1-stim7.

## 3. Discussion

In summary, lowering of F2, F3, and F4 seem to be acoustic cues for emotional (*sad*) speech, but breathiness (H1-H2 and H1-A3) does not seem to cue degrees in *sadness*. It is somewhat surprising that the measures of voice quality (H1-H2 and H1-A3) did not show a significant correlation with perception of emotion. One of our findings suggests that the stimulus that was well-perceived as emotional (stimulus 12) is less breathy than the one perceived as not emotional (stimulus 1), since often it is said that *sad* speech is breathy. This might not be so surprising as [7] has noted that there are there are different types of sad speech. Passive sad speech is usually described as having low F0, soft, and breathy, e.g., [8,9,10]. Active grieving speech, as is the case with the data in this study, is similar to what was reported for active grieving found in for instance, Russian laments[11]: F0 is high and there is a boost-up of energy around 4.5 khz, similar to that reported for the singer's formant [12]. Physiological and

modeling studies suggest that this boost-up of energy around 4-5 khz is caused by an expansion of the hypopharyngeal region, i.e., the area just above the vocal folds, e.g. [13]. The trough of energy around 5 khz, also seen in this data for the emotional speech, according to recent studies, may be caused by expansion of the side branches of the piriform fossa, a part of the hypopharyngeal region, e.g. [14,15].

About the finding of lowered F2, F3, and F4 for well-perceived emotional speech, we make the following comments. [16] showed that F2 INCREASED for emphasized /i/, as did F2-F1. But for emotional /i/, we saw the opposite: F2 decreases, as does F2-F1.However, for both emphasis and emotion, we see similar articulation patterns of jaw and tongue dorsum: jaw lowers and tongue dorsum raises and fronts. The reason given for this pattern of articulation for emphasis was that as the speaker opened the jaw lower for emphasis, she also raised the tongue dorsum up and forward, to make the high vowel more "diffuse," i.e., more "/i/-like." However, for emotional /i/, we suggest that F2 lowers because the larynx lowers, thus lengthening the vocal tract. Lowering the larynx would not only lengthen the vocal tract (thus lowering F2), it would also expand the hypopharyngeal region, both ventricular area and piriform fossa. The result would be lowering F3 and F4, increasing the energy around 4-5 kHz, and creating a trough around 5 khz. Lowering the larynx and expanding the hypopharyngeal region would produce a "sob" voice quality (see, e.g., Estill, 1992). Plans are underway to collect empirical data about larynx lowering during *sad* speech to check these hypotheses.

An interesting question would be why does the larynx lower? Why does the hypopharyngeal region expand when one is emotional (*sad*)? One possible answer may have to do with basic states of laryngeal settings--tensed vs. relaxed. For crying (as well as laughing), if the laryngeal setting is relaxed, perhaps this helps the speaker relax, which often happens after a bout of crying.

An interesting aside in terms of articulation of jaw and tongue position is that Erickson et al, (2006) found that the imitated emotional utterance (not spontaneous emotional utterance) showed no significant difference in jaw\tongue articulation from that of the imitated phrasing (no emotion) utterance, even though the imitated emotion was well-perceived (if not better perceived) as emotional, and the imitated phrasing (no emotion) was well-perceived as not emotional. A possible interpretation is that the way one changes articulatory setting to *convey* emotion/to communicate (*sadness*) to a listener, may be different than the way articulation changes if one is experiencing an emotion (*sadness*). We suggest this may be because emotion is a physiological/biological entity, and basic emotions may not have acoustic/articulatory targets, whereas acted emotions most likely do.

## Acknowledgements

We thank Masaaki Honda and NTT Communication Science Labs, Atsugi, Japan, for allowing us to use the EMA facilities to collect the original data, and Akinori Fujino for assisting in the data-collection. Also, we thank Albert Rilliard for helping with the software for the perception tests, and the college students at Capital University and Black Hills State University. This work was supported by the Japanese Ministry of Education, Science, Sport, and Culture, Grant-in-Aid for Scientific Research (C), (2007-2010):19520371 to the first author, and part of this work was supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan, and also by the Ministry of Education, Science, Sport and Culture, Grant-in-Aid for Scientific Research (A), 16202006, 19202013.

## 4. References

[1] Kawahara, H.; Matsui, H., 2003. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. *Proc. ICASSP, 2003*.

[2] Forsell, M., 2007. *Acoustic Correlates of Perceived Emotions in Speech*. Masters Thesis in Speech Communication, Royal Institute of Technology, KTH.

[3] ANSI, 1973. *Psychoacoustical terminology. Technical Report*, S.3.30, American National Standard Report.

[4] Menezes, C.**;** Maekawa, K.; Kawahara, H., 2006. Perception of voice quality in paralinguistic information types. In *Proceedings of the 20th General meeting of the Phonetic Society of Japan, Special issue of the 80th Anniversary*. Tokyo, Japan, 153-158.

[5] Estill, J., 1992. Primer of Compulsory Figures. Santa Rosa, CA: Estill Voice Training Systems.

[6] Erickson, D.; Yoshida, K.; Menezes, C.; Fujino, A.; Mochida, T.; and Shibuya, Y., 2006. Exploratory study of some acoustic and articulatory characteristics of *sad* speech. *Phonetica*, (63)1-25**.**

[7] Scherer, K. R., 1979. Nonlinguistic vocal indicators of emotion and psychopathology. In *Emotions in personality and psychopathology*, C. E. Izard (ed.).NY: Plenum Press, 493-529.

[8] Iida, A., 2000. *Study on Corpus-based Speech Synthesis with Emotion*, Doctoral thesis, Keio University, 2000.

[9] Eldred, S. H.; Price, D. B., 1958.The linguistic evaluation of feeling states in psychotherapy, *Psychiatry*,21,115-121.

[10] Mokhtari, P.; Campbell, N., 2003. Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech, in Sp. Issue on Speech Information Proc. *IEICE Transactions on Information and Systems*, (E-86-D), 574-582.

[11] Mazo, M.; Erickson, D.; Harvey, T., 1995. Emotion and expression, Temporal date on voice quality in Russian lament, *8th Vocal Fold Physiology Conference*,173-187.

[12] Sundberg, J., 1987. *The Science of the Singing Voice*. Dekalb, Ill.: Northern Illinois Univ. Press.

[13] Imagawa, H. ; Sakakibara, K. ; Tayama, N. ; Niimi, S., 2003. The effect of the hypopharyngeal and supra-glottic shapes for the singing voice. *Proc. Stockholm Musical Acoustics Conf. 2003*, (II), 471-474.

[14] Kitamura, T.; Honda, K.; Takemoto, H., 2004. Individual variation of the hypopharyngeal cavities and its acoustic effects. *Acoust. Sci. & Tech*

[15] Dang, J.; Honda, K., 1966. Acoustic characteristics of the piriform fossa in models and humans. *J. Acoust. Soc. Am.*, (101), 456-65.

[16] Erickson, D., 2002. Articulation of extreme formant patterns for emphasized vowels. *Phonetica*,(59), 134-149.