F-ratio of Speaker Variability in Emotional Speech

So Pae Yi*

ABSTRACT

Various acoustic features were extracted and analyzed to estimate the inter- and intra-speaker variability of emotional speech. Tokens of vowel /a/ from sentences spoken with different modes of emotion (sadness, neutral, happiness, fear and anger) were analyzed. All of the acoustic features (fundamental frequency, spectral slope, HNR, H1-A1 and formant frequency) indicated greater contribution to inter- than intra-speaker variability across all emotions. Each acoustic feature of speech signal showed a different degree of contribution to speaker discrimination in different emotional modes. Sadness and neutral indicated greater speaker discrimination than other emotional modes (happiness, fear, anger in descending order of F-ratio). In other words, the speaker specificity was better represented in sadness and neutral than in happiness, fear and anger with any of the acoustic features.

Keywords: Intra-speaker variability, Inter-speaker variability, F-ratio, Emotional speech

1. Introduction

Recent years have seen improvements in speech technology such as speech recognition, speaker recognition and speech synthesis. However, the performance of these technologies is influenced by many factors. Emotion is one of the sources that brings about performance degradation by inducing extra intra-speaker variability. It has been only rather recently that the characteristics of emotional speech have drawn the attention of phoneticians and engineers working on the various aspects of speech.

It is possible to think that the characterization of emotional speech would facilitate the speech classification thus compensating the negative impact from the speech in non-neutral modes. This paper explored the inter- and intra-speaker variability with different emotional modes by estimating the F-ratio of various acoustic features such as F0 (fundamental frequency), HNR (harmonic to noise ratio), Spectral Slope, H1-A1 (the amplitude difference between 1st harmonic component, 1st formant), F1 (1st formant frequency), F2 (2nd formant frequency), F3 (3rd formant frequency) and F4 (4th formant frequency).

* Cognitive Science, Pusan National Univ.

2. Method

2.1 Data

The emotional speeches produced by professional actors were analyzed. The actors were requested to speak ten sentences in six emotional modes (SiTEC, 2004). The recoding process was performed in a semi-sound proof room with an approximately -10dB shield effect (Jo et al., 2004). The sampling rate was 48000 Hz with 16 bit resolution. The recording equipment included AKG C414 B-ULS microphone and a sony 59ESJ DAT recoder. The emotional speeches explored in this study include neutral, happiness, anger, sadness and fear. The collected emotional data was evaluated by 20 judges using a subjective opinion test indicating an average consistency rate of 86.6% (SiTEC, 2004).

The vowel /a/ was segmented and extracted from each sentence using the voice analysis software PRAAT. The vowel /a/ was chosen because the 1st harmonic component and 1st formant are well separated in /a/ resulting in a reliable estimation of the 1st formant frequency and the amplitude difference between the 1st harmonic component and 1st formant (H1-A1). Ten tokens of vowel /a/ per each speaker (6 people) were extracted under each emotional mode (5 modes) resulting in 300 tokens (10 x 6 x 5). Acoustic features (F0, HNR, Spectral Slope, H1-A1 and formants) widely used to analyze speaker variability were estimated from the vowel segments. A window of length 20ms with a frame advance of length 10 ms was used for the feature estimation.

2.2 Analysis Procedure

In order to determine the degree of speaker discrimination under different emotional modes by an acoustic feature, the statistical F-ratio of inter- to intra-speaker variability was computed. F-ratio is a product of ANOVA and reflects both within- and between-sample variation (McCall, 2001). From a forensic speaker identification perspective, a higher F-ratio reflects greater inter- than intra-speaker variation; so the higher the F-ratio is, the more speaker-specific the parameter is (Wolf, 1972; Sambur, 1975; Rose, 2002).

The following was adapted from Khodai-Joopari's research.

The inter-speaker variability which is the variance of the speaker means weighted by the number of tokens per speaker was calculated via:

$$\sigma^2_{\text{int}er} = \frac{\sum_{i=1}^{N} n_i(\overline{X_i} - \overline{\overline{X}})^2}{N-1}$$

The intra-speaker variability which is the mean of the speakers' variances was calculated via:

$$\sigma^2_{\text{int}ra} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}(X_{ij} - \overline{X_i})^2}{\sum_{i=1}^{N} n_i - N}$$

Where n_i is the number of tokens (10) per speaker, N is the total number of speakers (6), X_i is a value of an acoustic feature;

$\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$ is the mean for ith speaker;

is the grand mean. $\overline{\overline{X}} = \frac{1}{N}\sum_{i=1}^{N}\overline{X_i}$

Mean values of each acoustic feature (F0, HNR, Spectral Slope and H1-A1) were computed from all 400 tokens. F0 was computed with autocorrelation method and HNR with cross-correlation method. Spectral Slope and H1 was estimated from LTAS (Long Term Average Spectrum) comparing the energy difference between the two frequency bands (0 to 1000 Hz vs. 1000 Hz to 4000 Hz). The F-ratio of each acoustic feature was obtained by dividing inter-speaker variability by intra-speaker variability.

3. Results and Discussion

As can be seen in <Figure 1> and <Table 1>, all acoustic features indicated greater contribution to inter- than intra-speaker variability (All of the F-ratio values were positive). Among all the acoustic features, sadness was found to achieve the greatest discrimination of speakers among all emotional modes followed by neutral, happiness, fear and anger in descending order.

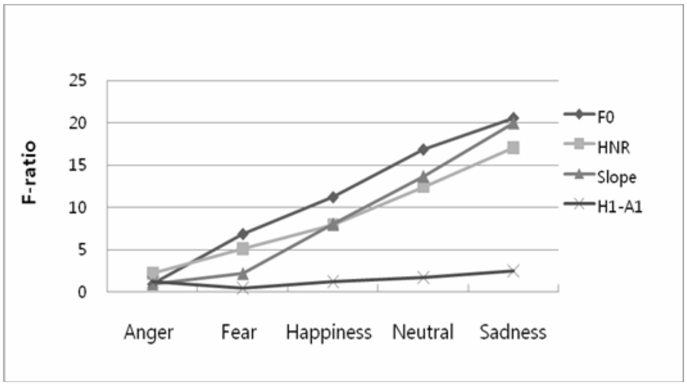

Figure 1. F-ratio of different acoustic features in different e motional modes

A possible explanation can be found in previous studies in one of which the mood of speakers while expressing the emotions of happiness, fear and anger was in a much higher arousal level than that of sadness (Cowie, 2001). Hence, the increased vocal effort led to higher degree of variability within an individual speaker. For example, frequency-dependant changes in level were found within an individual speaker when going from normal vocal effort (conversational speech) to increased vocal effort (shouted speech) that resulted in overall changes in the spectral tilt of the speech signal (Pickett et al., 1958). In another study, values of F0 were much less differentiated across speakers for increased vocal effort than for normal vocal effort (Rostolland, 1982).

<Figure 2> shows the LTAS of /a/ from each emotion explored in this study. It was observed that as the vocal effort increases, the spectral slope decreases. The noticeable difference between sadness and fear, as shown in Figure 2, was observed in the previous analysis where sad utterances had strong spectral damping in voiced segments whereas the utterances spoken with fear had very little spectral damping (Klasmeyer et al., 2000).

Another possible explanation of the high F-ratio phenomena in sadness and neutral can be found in a speaker verification study on emotional speech where the verification results for speech during sadness or neutral greatly outperformed those during happiness, fear or anger. This might be attributed to different levels of intra-speaker vocal variability when speakers are exhibiting different emotions (Wu et al., 2006). It also showed that when speakers are in the emotion of anger, fear or happiness, the pitch has a much wider range than that in sadness or neutral. This indicates that when speakers are in these three types of emotions (anger, fear or happiness), their articulating styles tend to create much greater intra-speaker vocal variability than they do in the emotion of sadness or neutral. So the articulating style of a certain type of emotions, which creates greater intra-speaker vocal variability, is one of the reasons for the performance decline of speaker verification on emotional speech (Wu et al., 2006).

Table 1. Intra and inter variability and F-ratio of different acoustic features in different emotional modes

F0	Anger	Fear	Happiness	Neutral	Sadness
Intra	20410.00	112768.27	259305.33	423491.29	615430.54
Inter	17929.26	768379.95	2896296.78	7119615.93	12634445.39
F-ratio	0.88	6.81	11.17	16.81	20.53
Spectral Slope	Anger	Fear	Happiness	Neutral	Sadness
Intra	57.22	337.15	782.12	1315.12	2093.56
Inter	52.44	727.49	6283.98	17935.91	41736.01
F-ratio	0.92	2.16	8.03	13.64	19.94
HNR	Anger	Fear	Happiness	Neutral	Sadness
Intra	69.22	315.18	779.51	1407.72	2236.49
Inter	150.30	1614.64	6148.05	17466.46	38140.71
F-ratio	2.17	5.12	7.89	12.41	17.05
H1-A1	Anger	Fear	Happiness	Neutral	Sadness
Intra	325.78	1239.24	2749.84	4599.71	6822.79
Inter	381.68	541.05	3260.23	7879.35	16938.79
F-ratio	1.17	0.44	1.19	1.71	2.48

The articulating style of sadness is shown to be distinguished from that of other emotional modes in the research (Jo et al., 2006) that analyzed the same emotional speech database as the one explored in this study. According to the research that compared each emotional speech with the neutral speech by computing the vocal tract ratio, sadness showed great difference in the lip section (fear also showed great difference for the lip section) and relative difference in the middle section whereas the other emotional modes indicated only a little difference in vocal tract ratio.

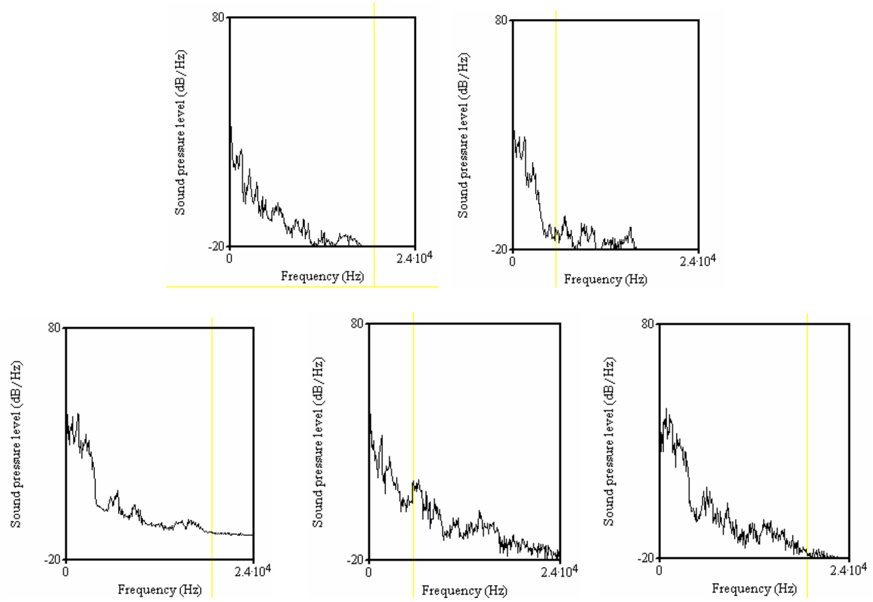

Figure 1. Long Term Average Spectrum of /a/ in sadness, neutral, happiness, fear and anger (from left to right and top to bottom)

Mean formant frequency values were computed from all 400 tokens and F-ratio was obtained from dividing inter-speaker variability by intra-speaker variability of each formant frequency (F1, F2, F3, F4). <Figure 3> and <Table 2> show that inter-speaker variability is consistently greater than intra-speaker variability of all formant frequencies (All of the F-ratio values were positive). With all formants, sadness was found to achieve greatest discrimination of speakers among all emotional modes followed by neutral, happiness, fear and anger in descending order.

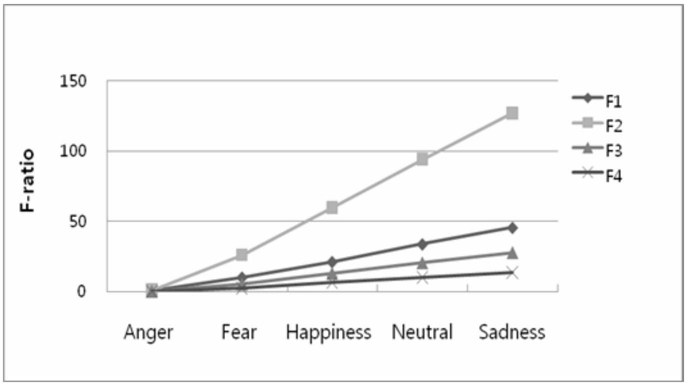

Table 2. Intra and inter variability and F-ratio of formant frequencies in different emotional modes

F1	Anger	Fear	Happiness	Neutral	Sadness
Intra	87801.08	936537.38	2342211.89	3939578.35	5932875.21
Inter	39470.06	9228849.42	49195453.38	132654029.01	269893978.47
F-ratio	0.45	9.85	21.00	33.67	45.49
F2					
Intra	232749.14	1294770.99	3194305.60	5667568.35	8740319.73
Inter	223475.15	33725065.00	190605458.11	532537070.36	1110179198.77
F-ratio	0.96	26.05	59.67	93.96	127.02
F3					
Intra	4319832.09	17354536.27	3194305.60	5667568.35	101523268.34
Inter	334599.80	90724828.38	484378091.53	1333516941.00	2786205500.98
F-ratio	0.08	5.23	12.75	20.22	27.44
F4					
Intra	14750762.84	62208860.89	140683315.51	5667568.35	401454860.63
Inter	475777.38	162591509.19	911276436.66	2521875015.04	5414191518.80
F-ratio	0.03	2.61	6.48	10.00	13.49

The F-ratio variation of formant values in different emotional modes in this study is consistent with the findings from previous researches in one of which the variations across emotional states in the spectral features were found at the phoneme level, especially vowel sounds (Leinonen et al., 1997). In another study, F2 was the most successful formant in speaker discrimination for /a/ and /i/ (Niessen, 2004). Furthermore, in a study of discrimination of speakers using the formant dynamic of /u:/ in British English, F2 analyses consistently provided higher levels of classification than F1 analyses.

In articulatory perspective, F2 corresponds to the position of the tongue in the mouth, i.e., front vs. back. Researchers investigating twins' speech have shown that identical physical dimensions do not necessarily give rise to identical articulatory behavior (Nolan et et al., 1996, Whiteside et al., 2003). Based on the study of speech produced by identical twin pairs, it can be said that formant variations across speakers were not simply caused by anatomical differences (Loakes, 2004). Considering the way the tongue impedes the vocal tract, it is likely that individual differences in configuration of the tongue during articulation is responsible for the greater inter-speaker variability.

4. Summary and Conclusion

Various acoustic features were extracted and analyzed from the vowel /a/ spoken with different modes of emotion (sadness, neutral, happiness, fear and anger). All of the acoustic features (fundamental frequency spectral slope, HNR, H1-A1 and formant frequency) indicated greater contribution to inter- than intra-speaker variability (All of the F-ratio values were positive). The different acoustic features of speech signal showed different degree of contribution to individual voice discrimination under different emotional modes. Sadness and neutral indicated greater F-ratio than other emotional modes (happiness, fear and anger in descending order). In other words, the speaker specificity was better represented by all of the acoustic features in sadness and neutral than in happiness, fear and anger.

Possible explanations for the preference of speaker discrimination under different emotional modes were discussed. One of them is the convergence of the speaker-dependent acoustic features at high levels of vocal effort leading to greater difficulty of speaker discrimination compared to the moderate level of vocal effort (e.g. conversational talking). Another explanation was presented with previous research showing that the different articulating styles of emotions create greater intra-speaker vocal variability.

Further research related to human discrimination of speakers in different emotional modes is needed. One of the studies showed that human subjects were so proficient at identifying the shouting talkers (Brungart et al., 2001) even though the acoustic features are indicating that speaker discrimination is more difficult at high levels of vocal effort than at moderate level. Other vowels such as /i/, /e/, /o/ and /u/ also should be explored.

References

Baayen, R., Piepenbrock, R. & Gulikers, L. 1995. *The CELEX Lexical Database*. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Boersma, P. & Weenink, D. 2008. Praat: doing phonetics by computer (Version 5.0.08) [Computer program], Retrieved Feb 1, 2008, from http://www.praat.org/.

Brungart, D., Scott, K. & Simpson, B. 2001. "The influence of vocal effort on human speaker identification", *EUROSPEECH-2001*, 747-750.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J. 2001. "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, Vol.18, no.1, 32-80.

Jo, C. & Bak, I. 2004. "Collection of Korean Emotional Speech Database from Actors", *Proceedings of ASK conference*, Vol. 23, No. 1, 45-48.

Jo, C. & Wang, J. 2006. "Determining the Relative Differences of Emotional Speech Using Vocal Tract Ratio", *Speech Sciences,* Vol. 13, No. 1, 109-116.

Khodai-Joopari, M., Clermont, F. & Barlow, M. 2004. "Speaker variability on a continuum of spectral sub-bands from 297-speakers' non-contemporaneous cepstra of Japanese vowels", *Proceedings of the 10th Australian International Conference on Speech Science and Technology,* Macquarie University, Sydney, 8 to 10 December, 504_9.

Klasmeyer, G. & Sendlmeier, W. 2000. "Voice and Emotional States - Chapter 15", *Voice Quality Measurement,* Singular Thomson Learning, San Diego, USA.

Leinonen, L. & Hiltunen, T. 1997. "Expression of emotionalmotivational connotations with a one-word utterance," *Journal of the Acoustical Society of America*, Vol. 102(3), 1853_1863.

Loakes, D. 2004. "Front Vowels as Speaker-Specific: Some Evidence from Australian English", *Proceedings of the 10th Australian International Conference on Speech Science & Technology,* Macquarie University, Sydney, 289-294.

McCall, R. B. 2001. *Fundamental Statistics for Behavioural Sciences* (8th edition) Wadsworth: California.

Niessen, M. 2004. *Speaker Specific Features in Vowels*, MS Thesis, University of Groningen Oude Kijk in ‘t Jatstraat, Groningen, The Netherlands.

Pickett, J. & Pollack, I. 1958. “Intelligiblity at high voice levels and the use of a megaphone,” *Journal of the Acoustical Society of America*, Vol. 30, 1100_1104.

Rose, P. 1999. “Differences and distinguishability in the acoustic characteristics of hello in voices of similar sounding speakers: a forensic phonetic investigation”, *Australian Review of Applied Linguistics*, Vol. 22, 1-42.

Rostolland, D. 1982. “Acoustic features of shouted voice,” *Acustica*, Vol. 50, 118_125.

Sambur, M. 1975. “Selection of Acoustic Features for Speaker Identification”, *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vol. 23(2), 176-182.

SiTEC. 2004. Korean Emotional Speech Database.

Whiteside, S. & Rixon, E. 2003. “Speech Characteristics of Monozygotic Twins and a Same-Sex Sibling: An Acoustic Case Study of Coarticulation Patterns in Read Speech” *Phonetica*, Vol. 60, 273-297.

Wolf, J. 1972. “Efficient Acoustic Parameters for Speaker Recognition”, *Journal of the Acoustical Society of America*, Vol. 51 (No.6, part 2), 2044-2056.

Wu, W., Zheng, T., Xu, M. & Bao, H. 2006. “Study on speaker verification on emotional speech”, *INTERSPEECH-2006*, paper 1124-Wed3CaP.7.

received: January 28, 2008
accepted: March 3, 2008

▲ So Pae Yi (Ph.D.)
Cognitive Science, Pusan National University
30 Changjundong, Keumjunggu, Pusan, 609-735, Korea
Tel: 010-5555-6305
E-mail: sopaeyi@pusan.ac.kr