A Study of Intra-Speaker and Inter-Speaker Affective Variability using Electroglottograph and Inverse Filtered Glottal Waveforms

Daniel Bone[1], *Samuel Kim*[1], *Sungbok Lee*[1,2] , *Shrikanth S. Narayanan*[1,2]

[1] Viterbi School of Engineering, University of Southern California, USA
[2] Department of Linguistics, University of Southern California, USA
dbone@usc.edu

Abstract

It is well-known that different speakers utilize their vocal instruments in diverse ways to express linguistic intention with some paralinguistic coloring such as emotional quality. The study of voice source features, which describe the action of the vocal folds, is important for a deeper understanding of emotion encoding in speech. In this study we investigate inter and intra-speaker differences in voicing activities as a function of emotion using electroglottography (EGG) and an inverse filtering technique. Results demonstrate that while voice quality features are good indicators of affective state, voice source descriptors vary in affective information across speakers. Glottal ratio measurements taken directly from the EGG signal are more reliable than measurements from the inverse-filtered glottal airflow signal, but the spectral harmonic amplitude differences of EGG are less useful than from inverse filtering.

Index Terms: emotion, speech production, voice source, EGG, inverse filtering

1. Introduction

In recent years, voice source features, which reflect the movement of the glottis, have drawn increased attention from speech research fields including those dealing with emotional speech. The motivation is due to the apparent connection between voice source activity and emotional expression.

Voice quality features have been examined in relation to emotional activity in [1],[2],[3]. Murphy examined glottal characteristics of different emotionally-styled voice types using an electroglottograph on a single speaker [1], determining that certain glottal quotients are useful in classifying emotion. Airas and Alku investigated a recently developed glottal flow parameter, normalized amplitude quotient, NAQ [2]. They demonstrated that the NAQ obtained after inverse filtering showed promise for the analysis of emotional content in continuous speech. Gobl and Chasaide synthesized different voice qualities for perception of affect experiments [3]. The results support the relationship between voice quality and affect, although they argue it is not a one-to-one mapping. Glottal ratios, i.e., open quotient and closed quotient, have also been examined for emotional speech. Moore created an automatic, closed-phase-based inverse filtering algorithm [4] for his study of clinically depressed speech [5].

Certain spectral features have been shown to be related to voice quality features. H1-H2 is the difference between the amplitude of the first harmonic, i.e., pitch, and the second harmonic. H1-H2 has been known to correspond to the open quotient, which is the fraction of the glottal period in which the vocal folds are open. H2-H4 has also been studied.

In this study, we extract features from both electroglottograph (EGG) and inverse filtered speech in order to analyze the expressive tendencies of speakers in the source domain. A better understanding of affective tendencies of the glottal activities across speakers can provide valuable information on individuality in emotional speech production as well as on the development of speaker adpatation algorithms for improved emotion detection and tracking in spontaneous speech.

Section 2 describes speeh data acquisition, parameter estimation, and analysis methods. Section 3 describes results of ANOVA and Fisher discriminant analysis, and a summary is given in Section 4.

2. Method

2.1. Data acquisition

Four untrained subjects (2 male and 2 female) were asked to read a set of sentences expressing 4 different emotions, *angry, happy, neutral,* and *sad*. In each emotional reading, they were also asked to read the sentences in three different loudness levels, *soft, normal,* and *loud*. The readings were repeated 5 times for each loudness level, producing a total of 60 instances of a single utterance. Changing loudness levels is known to generate confusion among emotions, and this is controlled in most experiments. However, the loudness variable is not considered as a factor in the current study.

During the reading, speech and EGG signals were simultaneously recorded using a 2-channel EG2-PCX by Glottal Enterprise, and transmitted to the computer through USB. The sampling rate of the recordings is 16 kHz and they are stored in 16-bit resolution.

In this work, we investigate the vowel /a/ in the word "father" because the source-filter interaction is minimized due to the relatively large lip opening. The phonemes were manually segmented after time alignment of speech and EGG via cross-correlation.

2.2. Measured Parameters

2.2.1. Electroglottography (EGG) measurements

EGG measures the change of contact area of the vocal folds during phonation. Two electrodes placed on the throat at the level of the larynx measure small changes in resistance caused by abduction and adduction of the vocal folds. EGG signals were recorded using a 2-channel EG2-PCX by Glottal Enterprise using two 34-mm diameter electrodes. After recording, the EGG signals were high-pass filtered (cut-off at 60 Hz) in order to remove the lower frequency component due to the vertical movements of the larynx during spontaneous phonation. After the high-pass filtering, these EGG signals were processed to measure a set of quotient-type parameters and spectral characteristics.

2.2.2. *Glottal Airflow Estimate*

An inverse filtering technique described in [4] was implemented to estimate the glottal airflow (GA). Closed-phase analysis utilizes the minimal glottal-vocal tract interaction that occurs when the vocal folds remain closed. This instant is difficult to estimate and may not occur in female speakers depending on phonation types. The technique uses many pitch-synchronous windows in disjoint regions. The windows are slid along the signal in order to create multiple estimates of the glottal waveform. The smoothest estimates are chosen as the glottal airflow estimate. In Figure 1 sample EGG signals (inverted) and corresponding glottal airflow estimates are plotted.

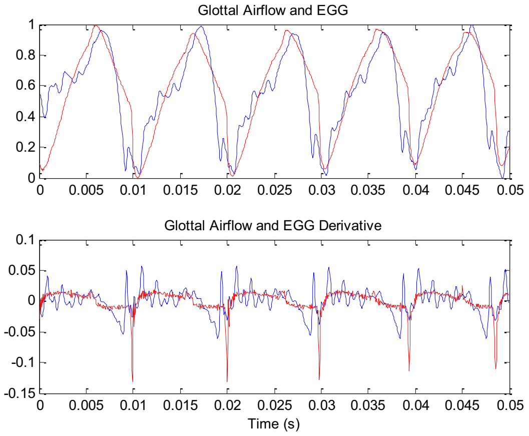

Figure 1: *Glottal airflow estimate (red) and inverted EGG (less noisy) in top figure and derivatives in bottom figure.*

2.2.3. *Glottal Quotients*

Ratios of opening phase to total pitch period, OQ, and ratios of closing to opening phases, rCPOP, were estimated, as well as the normalized amplitude quotient (NAQ).

EGG signals were first parameterized following [5] as shown in Figure 2. Green "*" represent maxima in the glottal waveform and red "x" represent minima. Magenta "o" represent minima in glottal derivative. Cyan "+" represent points of 15% rise in amplitude form minima to maxima. The minima point in the EGG derivative represents an approximation of glottal closure. The 15% rise in amplitude marks an estimation of the vocal folds opening.

In Figure 2, the open phase is represented by "O" and the closed phase by "C." The open phase is divided into opening period, Op, and closing period, Cl.

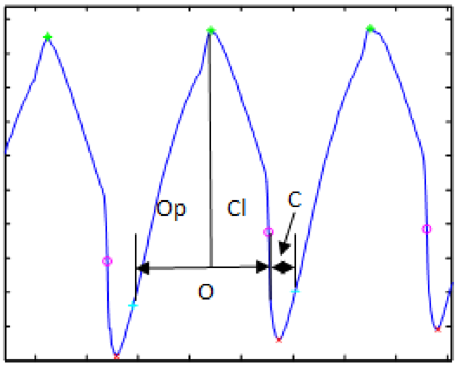

Figure 2. *Points of reference in glottal cycle set to estimate various ratio-type voicing parameters.*

The same quotient parameters were also estimated from the inverse filtered speech signals (i.e., glottal airflow estimates) after smoothing. Open quotient is calculated as the sum of open durations over the sum of total time for all cycles (taken from the minima for a particular segment). The ratio of closing phase to opening phase is calculated as the sum of closing times to the sum of opening times in a segment.

The normalized amplitude quotient, NAQ, is defined as peak to peak amplitude of the glottal airflow estimate divided by both the amplitude of the minimum of the derivative and by the period. The peak to peak amplitude is calculated as the amplitude of a maximum minus the amplitude of the next minimum. The minimum peak in the derivative is located between that previous maximum and next minimum. The equation is given in (1). f_{AC} is the peak to peak amplitude, d_{peak} is the peak in the derivative, and T is the period,

$$NAQ = f_{AC}/d_{peak}T \qquad (1)$$

2.2.4. *Harmonic Amplitude Differences*

Two types of harmonic amplitude differences, H1-H2 and H2-H4, are extracted using the tool VoiceSauce (downloadable at "http://www.ee.ucla.edu/~spapl/voicesauce/.") This program uses an algorithm developed in [6] to extract harmonic amplitude differences directly from speech by correcting for the effects of formants. The speech waveform is used as a reference during our analysis. Formant structures are non-existent in the EGG signals and largely eliminated in the glottal airflow estimates. Preliminary tests show that it is unreliable to correct the residual formant structures of the glottal airflow estimates. Therefore, for EGG and glottal airflows, uncorrected outputs are computed using the VoiceSauce software.

2.3. Statistical analysis of measurements

Results are presented for interpretation based on two statistical analysis methods: one-way ANOVA with accompanying means plots and Fisher linear discriminant analysis. Focus is given to investigating the effects of emotion expression on the voice source parameters and inter-speaker differences. SPSS was used for all the statistical analyses performed in the study. It is noted that all ANOVA results are reported at the 0.05 significance level. It is also noted inter-speaker analysis is performed on features that are z-normalized per-speaker.

Due to the minimal length segment required to get output from VoiceSauce and the failure to produce a periodic GA or EGG signal in some cases, only 198 of the original 240 utterances were kept. The following is a list of the number of utterances per speaker; M1: 56, M2: 37, F1: 59, F2: 46

3. Results and Discussion

3.1. ANOVA

Figures 3-6 display means plots with standard deviation bars for the various measures. As can be observed in Figure 3, for H1-H2 measured from GA, there is a difference between angry and happy for each speaker, and the difference is significant for M1 and F1. There also exist significant inter-speaker differences across emotions. For H1-H2 measured from EGG in Figure 4, however, there is only a significant difference for speaker M1 where the sad mean differs from angry and happy means. Regarding inter-speaker differences, the mean of angry differs from happy and sad across subjects.

In Figure 5, where NAQ was estimated from the inverse filtered speech signals, the observable pattern between speakers is that means for angry are lower than means for happy. Speakers F1 and F2 show significant differences between angry and all other emotions, while M1 and M2 show significance between all emotions. Across speakers, mean for angry is significantly distinct from all other emotions' means. It is interesting to observe that female GA NAQs are noticeably higher than male NAQs across all emotions.

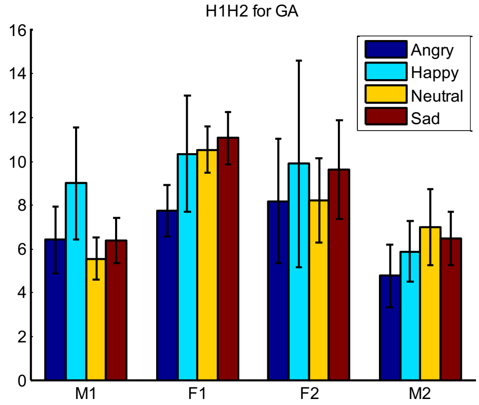

Figure 3. *Glottal airflow (GA) H1-H2 means plots.*

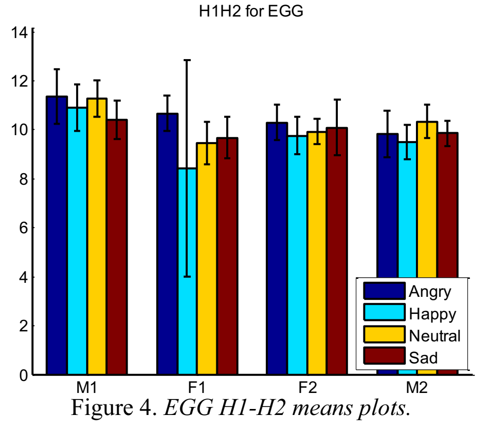

Figure 4. *EGG H1-H2 means plots.*

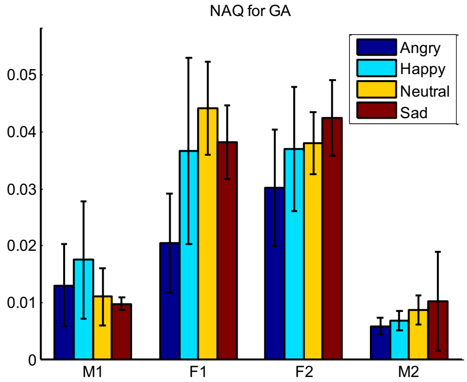

Figure 5. *Glottal airflow (GA) NAQ means plots.*

In Figure 6, NAQs estimated from EGG signals are plotted. Like GA, female EGG NAQs are normally higher than male NAQs across all emotions. ANOVA shows no significant differences in means for M1, but the other speakers show an assortment of differences between various emotion pairs. Across speakers, angry mean is different from happy mean.

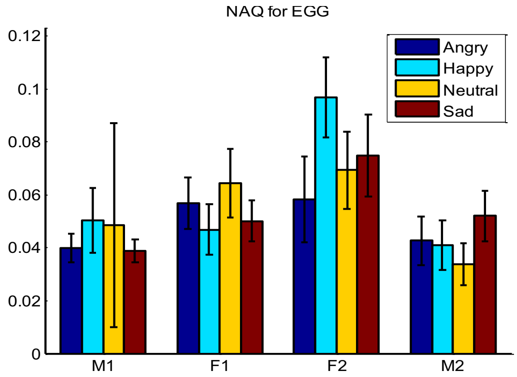

Figure 6. *EGG NAQ means plots.*

ANOVA may suggest that separation between emotions is possible for H1-H2 from inverse filtered glottal airflow (angry/happy) but not as possible for H1-H2 from EGG. This may be because the mapping from vocal fold activity (EGG) to glottal airflow seems non-linear and complex.

This analysis also may indicate that NAQ from the glottal airflow estimate and from EGG will aid in separation of emotions (except EGG-M1).

3.2. Fisher Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is conducted for three sets of features and the classification accuracy (%) is listed in Table 1. Harmonic amplitude differences (HADs) and glottal quotients from inverse filtered glottal airflow estimates (GA), EGG, and speech are used to discriminate emotions. For "All" features, data was z-normalized per-speaker and combined.

Table 1: *LDA results for individual source features.*

	M1	F1	F2	M2	M's	F's	All
H1H2ga	34	37	22	51	41	30	34
H2H4ga	38	37	27	46	41	33	36
H1H2egg	38	26	34	24	32	30	35
H2H4egg	14	44	51	54	30	47	30
H1H2sp	27	28	27	51	37	28	39
H2H4sp	45	20	19	43	44	19	27
GA:OQ	46	46	37	24	38	42	40
rCPOP	45	37	39	54	48	38	36
NAQ	34	41	37	41	37	39	26
EGG:OQ	39	9	41	30	35	23	29
rCPOP	38	20	41	46	41	29	32
NAQ	38	52	44	49	42	49	30

There are differences in average performance of features for males and females. The glottal airflow (estimate) H1H2 for males is better (emotion classification rate was higher) than EGG H1H2. The EGG H2H4 is better for females than the glottal airflow H2H4. HADs from speech classified best for males and poorly for females. Results suggest classification may obtain larger gain from speech-derived HADs for males.

Since EGG directly measures the activity of the glottis, it is expected that EGG timing quotients will be more accurate than inverse filtered glottal airflow features. EGG NAQ is better than GA NAQ for each speaker and across speakers. But, GA OQ and rCPOP predict emotion better on average and across speakers than EGG. This may happen because these quotients are only estimates of the actual phases.

It may be interesting to consider the top performing features for sets of speakers. The top three features for males are GA rCPOP, speech H2H4, and EGG NAQ. The top four features for females are EGG NAQ, EGG H2H4, GA OQ, and GA NAQ. The top features across all speakers (normalized) are GA OQ, GA rCPOP, GA H2H4, and EGG H1H2. The inverse filtered OQ and rCPOP, along with EGG NAQ and H2H4, seem to predict emotions well in many cases.

Table 2: *LDA results for combined source feature sets*.

	GA Set1	EGG Set1	Sp Set1	GA Set2	EGG Set2	GA Set3	EGG Set3
M1	32.1	25	42.9	46.4	51.8	37.5	58.9
F1	56.5	41.3	41.3	47.8	43.5	54.3	56.5
F2	37.3	49.2	40.7	45.8	54.2	49.2	54.2
M2	51.4	37.8	54.1	48.6	56.8	48.6	73
M’s	39.8	30.1	47.4	47.3	53.8	41.9	64.5
F’s	48.1	44.8	41.0	46.9	48.2	52.1	55.5
All	32.1	25	39.4	45.5	30.3	43.4	33.3

Three sets of features are considered in Table 2:

Set 1: H1H2, H2H4.
Set 2: OQ, rCPOP, NAQ.
Set 3: Set1+Set2.

It may first be noted for a given set, any one signal may not best capture the affective information across all speakers. Speech HADs are the best predictors of emotion twice, GA HADs are best once, and EGG HADs are best once. For male speakers, speech HADs are best, followed by GA HADs. GA Qs classify better than Speech HADs three times. This may be further evidence that extracting source information without inverse filtering is not enough (although H1A1 and similar features were not included in this analysis). The top performing set for all speakers is EGG Set 3 (all glottal features). The only place this set is not superior is in generalizability to inter-speaker classification.

GA set 3 predicted emotion much better for female speakers. This is interesting because the oppositee trend is seen for EGG set 3. It is noted that GA set 3 performs worse for the male speakers than GA set 2. Inverse filtering is usually more effective for males than females because of the longer closed period. The decrease in performance may suggest some redundant information between sets 1 and 2 for male speakers.

Correlation analysis of speaker M1 shows correlations between H1H2 and OQ of -0.721, between H1H2 and rCPOP of 0.882, between H2H4 and OQ of 0.868, between H2H2 and rCPOP of -0.810 at 0.01 significance level. For speaker M2, correlations exists between H1H2 and OQ of -0.697, and between H1H2 and rCPOP of 0.850 at the 0.01 significance level. Speaker F1 shows a correlation between H1H2 and NAQ, while speaker F2 shows correlations between H1H2 and rCPOP, and H2H4 and OQ of 0.777, 0.731, and 0.730 respectively at the 0.01 significance level. This may suggest that features from set 1 and set 2 were more correlated for males and created over-reliance on a certain feature that did not generalize well during classification.

When comparing the sets of GA features that classified best by gender, males obtained the best performance from set 2, lost some performance when adding set 1 to obtain set 3, and the worst classification rate was by set 1. Set 3 classified emotions best for females, followed by set 1.

The most generalizable set of features is GA set 2. This is interesting since its overall accuracy was not the highest, indicating that strong trends exist among features in this set. GA OQ had a performance of 40% versus 45.5% for set 2.

4. Summary

In this study, we investigated inter and intra-speaker differences in voicing activities as a function of emotion using electroglottography (EGG) and an inverse filtering technique.

Results suggest that Fourier harmonic amplitude differences (H1-H2 and H2-H4) from EGG do not contain as much emotional content as H1-H2 and H2-H4 from speech or inverse filtering. However, temporal domain features extracted from the EGG waveform are shown to better predict emotions as a group than those features from inverse filtering. It seems that for NAQ, EGG is more consistent than inverse filtering. For OQ and rCPOP, inverse filtering predicts better, possibly indicating more accurate estimates of cycle boundaries. Interspeaker results indicate that EGG features do not generalize well, but glottal ratios from inverse filtering do.

The combination of spectral and temporal features from inverse filtering provided an improvement in classification accuracy only for females. The strong correlation between those two sets of features for males may have caused the decrease in performance when the sets were combined.

It was seen that EGG spectral and temporal features outperformed all other tested sets of features for each speaker. The emotional content of the EGG signal is argued to be high, but generalized trends across speakers were not seen, and this may explain the poor overall inter-speaker accuracy. Glottal temporal ratios alone tended to outperform harmonic amplitude differences extracted directly from speech.

In the future, those aforementioned inter-speaker and inter-gender characteristics of voice source parameters will be explored toward the development of speaker adaptation methods for a better emotion recognition performance. The methods used may be extended to a “real” emotional database such as call-center data and other classifiers may be tested.

5. Acknowledgements

Work was supported by NSF, NIH, and USC Annenberg Fellowship.

6. References

[1] Murphy, P.J., Laukkanen, A.-M. “Electroglottogram analysis of emotionally styled phonation.” in *Multimodal Signals: Cognitive and Algorithmic Issues*., vol. 5398. Esposito, A., Hussain, A., Marinaro, M., Martone, R., Eds. Heidelberg: Springer, 2009, pp. 264–270.

[2] Airas, M., Alku, P. “Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalized amplitude quotient.” *Phonetica*, vol. 63, pp. 26–46, 2006.

[3] C. Gobl, A. N. Chasaide. “The role of voice quality in communicating emotion, mood and attitude.” *Speech Communication*, vol. 40, pp. 189-212, 2002.

[4] Moore, E.; Clements, M.A.; Peifer, J.W.; Weisser, L. "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech." *Biomedical Engineering, IEEE Transactions on* , vol.55, no.1, pp.96-107, Jan. 2008.

[5] E. Moore, II and M. Clements. “Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information.” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 101–104, 2004.

[6] Iseli, M., and A. Alwan. “An Improved Correction Formula for the Estimation of Harmonic Magnitudes and Its Application to Open Quotient Estimation.” in *Proc. ICASSP*, pp. 669-672 , Montreal, May 2004.