Continuous Spoken Emotion Recognition Based on Time-Frequency Features of the Glottal Pulse Signal within Stressed Vowels

Li Tian, Catherine Inez Watson

Department of Electrical and Computer Engineering, University of Auckland, New Zealand

tli725@aucklanduni.ac.nz, c.watson@auckland.ac.nz

Abstract

In speech production, emotional cues can be detected via three main aspects: excitation source, vocal tract and prosodic pattern. This paper addressed the first one, extracting six time and frequency related features from glottal pulse signals, transformed from stressed vowels. Four sustained vowels incorporating five regular emotions, which were selected from sentence recordings of the Berlin emotional speech database were investigated. The effectiveness of those glottal pulse features towards emotion recognition was proven through double round Robin quadratic classification in both single-gender and cross-gender stages, reaching average overall hit-rate of 63%, 64% and 53% for male, female and cross-gender respectively.

Key words: IAIF, glottal pulse, open quotient, speed quotient, frequency tilt, double round Robin Classification

1. Introduction

Most current human-machine speech communication systems are implemented with a simple one-channel interaction which merely transmits explicit verbal messages but lacks the capability of digging out the hidden intent, motive, and physiology state clues which speech may convey in the emotion of the speakers. Spoken information exchange cannot be fully achieved with only semantics. Some subtle acoustic characteristics embedded in emotion should also be captured to express intentions during speech synthesis.

According to the source-filter speech production model [1], speech can be viewed as the convolution of the glottal source, vocal tract filter, nasal cavity, lip radiation and articulation noise. Emotional clues have been extensively observed in various spectral features derived from the vocal tract (see [2] for an excellent summary of the studies). Several studies have also investigated some glottal-based features that are capable of classifying speech stress and emotion-related health confusions [3-5]. However, there is still not enough evidence compared with the vocal tract to address the contribution of the glottal source related features in differentiating emotion states. This is the main focus of this paper.

Iterative adaptive inverse filtering, IAIF [6] was applied in this study to derive glottal pulse signals from an emotional database. IAIF is a two-stage iteration process based on the principle of discrete all pole modelling (DAP) which recursively estimates the vocal tract model for every Hanning windowed 25ms analysis frame across every vowel. The glottal pulse signal is obtained by inverse filtering the vocal tract and lip radiation models from the original speech.

In this study six glottal pulse descriptive features are investigated in Section 2. Three are time-domain features, and they are mean open quotient (mOQ), standard deviation open quotient (stdOQ) and mean speed quotient (mSQ). Three are frequency-domain features, which are all methods of representing the spectral tilt. In Section 3 Principle Component analysis is used to reduce these six features to four dimension variables, which are combined as input vectors for subsequent classification model construction. Double round Robin classification [7] rather than the conventional support vector machines [8] was adopted in classification stages to examine the recognition performance, these results are discussed in Section 4. Conclusions are given in Section 5.

2. Extraction of glottal pulse features

2.1. Stressed vowel selection

The speech corpus used in this study is the Berlin emotional speech database Emo-DB [9] collected from ten (five male and five female) professional actors. Utterances were randomly sorted and based on ten linguistically neutral German sentences. Recording were stored with a 16 kHz sampling frequency as 16 bit numbers. There were seven emotions in the corpus: anger, boredom, disgust, fear, happiness, neutral and sadness, and there was an unequal distribution of utterances for each emotion. In this corpus the emotion of disgust achieved the worst subjective listening recognition rate (79.6%) [9] and the level of fear significantly differed from speaker to speaker. Therefore, for reliability and consistency of the recognition result, only exemplars of the remaining-five emotions were investigated in this study, yielding 725 sentences. Those emotional sentences were automatically labelled at word and phonetic level by the Munich Automatic Web Segmentation System, webMAUS [10]. All labelled sentences were converted into an EMU formatted database [11] and each vowel's onset and offset were extracted. To minimize potential errors in the detection of vowel boundaries, the first and last 5% of those extracted vowels would be removed before further processing.

In order to successfully recognize emotions from the glottal signal, we found that the length of selected vowel tokens must be over 65ms. In continuous speech only long tense monophthongs, such as **a**: can fulfil this constraint. Only four common sustained vowels /**a**: **o**: **i**: **e**:/ were examined in this study. It is natural that not all the extracted vowel tokens contain sufficiently distinguishable emotional clues. It is suggested that only those stressed vowels are able to clearly identify the emotional states of their corresponded sentence. To automatically identify stressed vowels three prosodic characteristics were examined for each vowel: (a) long duration, (b) changing pitch, (c) strong intensity. Any vowel token possessing at least two of above characteristics was deemed to be stressed. Thresholds were used to determine whether the three prosodic characteristics were present, however stressing thresholds altered with different emotions. For sadness and boredom the three thresholds were 80ms, 10Hz and 70rms respectively. For happiness and neutral the thresholds were 80ms, 10Hz, 75rms while for anger the values

were 90ms, 15Hz, 80rms. The reason why higher threshold values were used for anger was to lower the number of selected vowel tokens. If this was not applied there would be a huge bias towards anger in the analysis dataset. Table 1 summarize the total number of available vowels and final selected stressed ones for each emotion and vowel type.

Table 1. *Available emotional vowels and identified stressed vowels (in parentheses) in Emo-DB*

	Angry	Sad	Neutral	Happy	Bored
e:	154(43)	68(43)	85(44)	81(48)	93(47)
a:	104(40)	60(40)	59(41)	54(39)	59(35)
i:	149(37)	68(40)	97(45)	74(33)	89(50)
o:	53(18)	29(22)	32(22)	27(18)	34(20)
sum:	460(138)	225(145)	273(152)	236(138)	275(152)

2.2. Time domain features of the Glottal Pulse

To properly parameterize the shape of glottal pulse waveform, three types of feature points must be accurately identified.

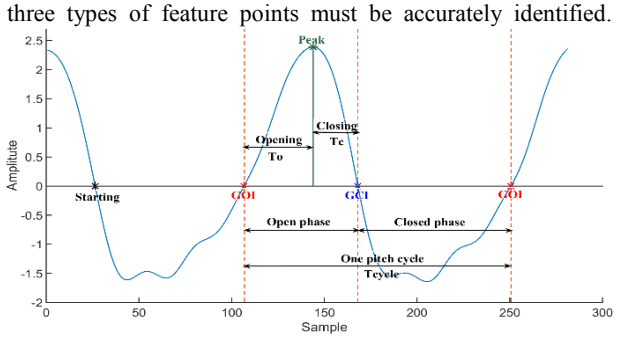

Figure 1: *Glottal pulse wave phases*

These are the glottal opening instance (GOI), the glottal closing instance (GCI) (both with an amplitude of 0), and the peak point between each GOI and GCI, indicating the maximal glottal opening when the glottal air flow at a maximum (see in Figure 1). The duration of two neighboring GCIs or GOIs represents one pitch cycle and the duration between each GOI and GCI pair can be regarded as the glottal open phase. In this study IAIF method has shown to give robust results of locating GOI and GCI independent of speakers, genders and emotions. Due to the sampling quantization limit that each two consecutive sample points have constant time spacing 0.0625ms, GCI and GOI points rarely coincide with sampled points, therefore linear interpolation between each two zero-crossing adjacent points is used for more precise GCI and GOI detection [12].

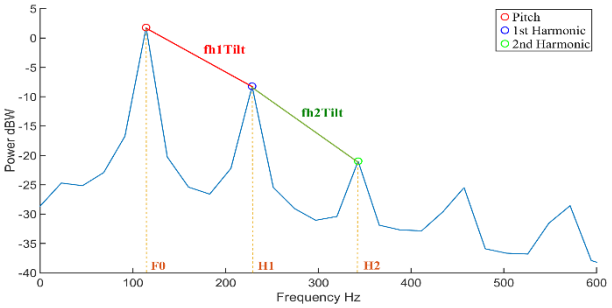

Figure 2: *Glottal pulse spectral tilt*

In order to derive the time domain features from the glottal signal it is important to identify the starting point first, which is the zero crossing prior to the first GOI (see Figure 1). The starting point is the very first detected GCI point rather than GOI since GCI theoretically correspond to the linear prediction residual peaks which are easier to find [13].The open quotient is the time ratio of when the glottal folds are open to the corresponding span of the pitch cycle and the speed quotient is given as the ratio of the opening phase over the closing phase.

$$OQ = \frac{T_o + T_c}{T_{cycle}} \quad (2)$$

$$SQ = \frac{T_o}{T_c} \quad (3)$$

To derive the glottal pulse and calculate the open quotient and speed quotient from each vowel segment, a sliding rectangle window of fixed length with a 1ms shift was used. The OQ is calculated in a portion of the waveform which is deemed to be stable. The analysis window moves until stability is reached, where the mean OQ values from the present and previous two windowed segments vary within a reasonably small margin. The window length and variance margin are adjusted according to input vowel length (see in Table 2), these values were determined heuristically. Typically there are between 4 to 16 glottal cycles in an analysis window. Results from the final window interval are collected to calculate the wanted mOQ, stdOQ, mSQ values.

Table 2. *Preset values for different length vowels*

Vowel length(ms)	Window length(ms)	Variance margin
(65,100]	45	0.02
(100,150]	60	0.01
(150,250]	80	0.006

2.3. Frequency domain feature

When converting signal from time domain to frequency domain, there is always a tradeoff between frequency resolution and noise-reduction [14]. In this study the main concern is noise, which has been introduced by both irregular glottal closures and extremely short sample lengths. Various frequency tilt values in the region between 0 to 3700 Hz have been investigated and successfully used in depression disorder identification [8]. Inspired by this, two different spectral tilts will be examined for emotion discrimination. Figure 2 gives an example of Welch transformed frequency plot of a glottal pulse signal converted from a /**a**: / token in the database. The spectral peaks can be easily identified and two lines can be well fitted to the first three peaks. Three frequency features depending on these peaks are defined as:

$$fh1Tilt = \frac{P(f_0) - P(h_1)}{f_0 - h_1} \quad (4)$$

$$fh2Tilt = \frac{P(h_1) - P(h_2)}{h_1 - h_2} \quad (5)$$

$$hf = \frac{P(f > h_2)}{P(f > 0)} \quad (6)$$

Where $P(f_0)$ represents the power at fundamental frequency, $P(h_1)$ and $P(h_2)$ are the power of first and second harmonics. The features fh1Tilt and fh2Tilt describe the two-step power dropping rate from the fundamental frequency to the second harmonic while hf specifies the cumulative power impact of those high frequency (>h_2) components.

3. Classification design

Some of the six extracted features were highly correlated, and therefore there would be redundant information if all six features were used in a classifier. Moreover too much redundancies can cause an over-fitting problem which inevitably decreases the classifier performance. Thus principle component analysis (PCA) transformed the data such that the variability of the data was accounted for by fewer (rotated) features than the original set. A Shapiro-Wilk normality tests

[15] established that a normal Gaussian model would suffice for describing the subset data distribution for each emotion and vowel type in the classifier. The average Shapiro results and overall importance of the 6 principle components is given in Table 3. The separability of five emotions can be seen in Figure 3, a scatter plot of the first two PCA components for monophthong **a**:. Different colored circles indicate different emotional entries. It can be observed that even with only two dimension the 5 emotions are well separated, there are not substantial overlaps among different clusters. In particular boredom is quite separate from anger and happiness. Similar finding were observed for the other vowels.

Table 3. *Averaged Shapiro result and overall importance of PCA components monophthongs*

	PC1	PC2	PC3	PC4	PC5	PC6
Importance	31%	25%	18%	14%	7%	5%
Shapiro w	0.47	0.43	0.19	0.22	0.52	0.45
Shapiro p	0.96	0.97	0.95	0.97	0.98	0.97

The ideal number of principle components to compactly characterize the six features is constrained by the open-test recognition performances. Multiple experiments with respect to all four sustained monophthongs demonstrated that the over-fitting problem negatively impacted on the cross-gender open-test performances when using the first five PCA components. Therefore recognition training and testing was carried out on the first four PCA components which contributed 88% of the dataset variability.

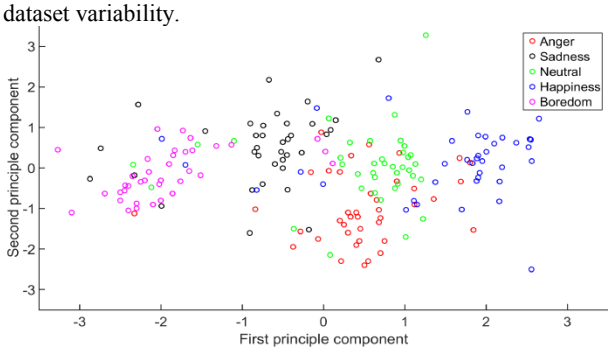

Figure 3: *PCA first two components scatter plot*

In this study, classification was conducted in three stages for different monophthongs. The first two stages were designed to classify emotions within same gender, i.e., the testing data and training data came from the same gender domain, either male or female, while the third stage was a cross-gender open-test where the testing and training data were from female and male datasets respectively (both of which had a similar sample size). A double round Robin quadratic Gaussian classification was deployed throughout.

The double round Robin classification is more powerful than those typical one-vs-all classifiers on the restricted training multi-class classification problem [16]. In this study, it turned the five-class problem into twenty binary problems, one for classifying each pair of emotional classes. For each binary problem, a base learner which was a four dimension quadratic Gaussian classifier decided to which of its two classes an input testing sample was more likely to belong. The winning emotion would be assign a "ticket". After twenty judgments, the emotion which had accumulated the most tickets would be the final prediction of the testing entry. Possible ties would be broken by preferring the emotion with more training samples. All the testing samples would go through this procedure one by one and a confusion matrix would be formed at the end of the process.

Furthermore in order to investigate whether the low frequency speech components contributed most of emotion separation, the classification process was repeated on features calculated from low-pass filtered vowel tokens. The low-pass filter was a FIR filter with a cutoff frequency of 1000Hz.

4. Result and Discussion

Recognition confusion matrices for four monophthongs under male, female and cross-gender stages are given in Table 4. Instead of the number of samples, the percentages of correctly identified tokens, for each emotion are shown. Overall hit-rates (the sum of diagonal elements divided by the entire sum in each matrix) across all categories are given in Table 5 together with the results from the filtered data.

Results show that anger is obviously the best recognized emotion regardless of the vowel types and genders with an average hit-rate of 75% followed by boredom and neutral, although these had varying performances with respect to different categories (male **o:** the best and cross-gender **i**: the worst). In some cases, such as the cross-gender test for **i**:, sadness tokens were substantially misclassified as boredom. For the most part happiness often confused with anger, especially for the monophthong **a**: and **o**: which could be expected when looking at Figure 3. This was partly because both anger and happiness have a short intense open phase and a long relaxed closed phase. Thus it was difficult to separate those time domain features for them, however the differences from their frequency-domain could still provide meaningful classification clues. The relatively more abrupt glottal air puffs when speaking in anger raised the power of first and second harmonics in the frequency plot, resulting in less steep spectral tilt than that of happiness. This can been seen in analysis of efficacy of the time and frequency domain features in Table 6. The frequency features alone can provide much better anger and happiness recognition than the time features alone and in the case of **a**: and **o**: the recognition performances of the combination of six features even cannot compete with frequency features alone. There was also a difference between the vowels in the emotion recognition scores with rates for the back vowels **a**: and **o**: considerably better than those for the front vowel **i**: and **e**:. Male and female had almost the same average overall recognition hit-rate, 64% and 63% respectively, when the training and testing sets were of the same gender. In the cross-gender scenario (different genders for the training and testing sets) the recognition performance decreased sharply to around 50%. This is possibly due to gender differences in the glottal pulse features coupled with the insufficient cross-gender training which is bounded by the size of the applied corpus. Ideally more extensive data especially the training data should be required.

Table 5. *Overall hit-rate (%) table for raw monophthongs (bold) and filtered ones under three different stages.*

	a:|a:	**e**:|e:	**i:**|i:	**o:**|o:	Average
Male	**67**|66	**61**|59	**61**|58	**62**|60	**63**|61
Female	**74**|67	**58**|53	**56**|50	**67**|61	**64**|58
Cross-gender	**53**|44	**52**|46	**50**|43	**56**|48	**53**|45
Average	**65**|59	**57**|53	**56**|50	**62**|56	

5. Conclusions

This work highlighted the capability of the IAIF transformed glottal pulse signals in emotion recognition using their six time and frequency related features. The whole process was free of

Table 4. *Emotion recognition hit-rate (%) confusion matrices for different monophthongs under three different stages.*

		Male					Female					Cross-gender				
		Angry	Bored	Happy	Neutral	Sad	Angry	Bored	Happy	Neutral	Sad	Angry	Bored	Happy	Neutral	Sad
a:	Angry	**90.0**	0	10.0	0	**0**	**85.0**	5.0	10.0	0	0	**75.0**	0	25.0	0	0
	Bored	0	**76.5**	5.9	0	17.6	0	**52.4**	4.8	19.0	9.5	16.7	**44.4**	11.1	22.2	5.6
	Happy	37.5	6.3	**37.5**	12.5	6.3	21.7	0	**69.6**	8.7	0	56.5	4.3	**21.7**	13.0	4.5
	Neutral	0	25.0	0	**55.0**	20.0	0	19.0	9.5	**66.7**	4.8	4.8	4.8	14.3	**61.9**	14.3
	Sad	0	16.7	0	11.1	**72.2**	0	4.5	0	9.1	**86.4**	0	0	31.8	0	**68.2**
e:	Angry	**72.2**	16.7	5.6	0	5.6	**74.0**	7.4	14.8	0	3.7	**59.3**	0	22.2	0	18.5
	Bored	0	**54.2**	0	41.7	4.2	0	**73.9**	0	21.7	4.3	0	**47.8**	13.0	17.4	21.7
	Happy	5.3	10.5	**52.6**	21.1	10.5	20.7	31.0	**41.4**	6.9	0	19.4	9.7	**38.7**	6.5	25.8
	Neutral	0	13.6	9.1	**72.7**	4.5	0	36.4	0	**63.6**	0	0	36.4	9.1	**45.5**	9.1
	Sad	4.5	13.6	4.5	27.3	**50.0**	9.5	33.3	4.3	4.8	**48.1**	4.8	14.3	0	4.8	**76.2**
i:	Angry	**70.6**	11.8	11.8	0	5.9	**90.0**	5.0	5.0	0	0	**80.0**	0	15.0	5.0	0
	Bored	0	**70.8**	0	4.2	25.0	3.8	**61.5**	3.8	7.7	23.1	11.5	**42.3**	15.4	23.1	7.7
	Happy	11.1	22.2	**38.9**	16.7	11.1	33.3	6.7	**33.3**	13.3	13.3	33.3	13.3	**46.7**	0	6.7
	Neutral	0	45.5	0	**50.0**	4.5	8.7	43.5	0	**21.7**	26.1	8.7	13.0	17.4	**60.9**	0
	Sad	0	11.8	0	11.8	**76.5**	8.7	21.7	0	0	**69.6**	4.3	43.5	21.7	8.7	**21.7**
o:	Angry	**71.4**	0	28.6	0	0	**72.7**	0	18.2	9.1	0	**81.8**	9.1	0	9.1	0
	Bored	0	**75.0**	0	0	25.0	0	**50.0**	0	37.5	12.5	0	**47.5**	12.5	27.5	12.5
	Happy	60.0	0	**30.0**	0	10.0	40.0	0	**60.0**	0	0	58.0	0	**42.0**	10.0	0
	Neutral	0	12.5	12.5	**75.0**	0	8.3	16.7	0	**58.3**	16.7	0	25.0	25.0	**50.0**	0
	Sad	0	25.0	0	12.5	**62.5**	0	0	0	14.3	**85.7**	7.1	14.3	14.3	21.4	**42.9**

manual involvement and computational hassle. Theoretically it could extend to any emotional corpus of any size with only requirements of sentence recordings and corresponding contexts. Original and low-pass filtered speech signal were both investigated. Results indicated that male and female data had decent recognition performances with average overall hit-rates of 64% and 63% respectively and the gender differences may negatively impact on the recognition performances when database were mixed with dual gender speakers such that for a more sophisticated judgement, two hierarchy structure should be built where gender identification should be on the top of emotion classification. Besides, most glottal characteristics associated with emotion discrimination was confirmed to come from the low frequency portions of the speech especially for male speakers.

In future, a more balanced and expanding corpus should be created so as to reinforce the outcome's reliability and minimize the emotion's variability within speakers and sentences. The weights of features extracted from time domain and frequency domain should be treated differently for some specific pairs of emotions like anger and happiness to increase correct recognition ratio. The capture of the nonlinear relations present in the glottal features with more complex non-linear model may also benefit to the recognition performances. Practically the fusion of linguistic and prosody characteristics will complement to glottal pulse features and as a whole push the emotion recognition to a more robust level.

Table 6. *Two-class (anger and happiness) average overall hit-rates (%) for different features and monophthongs.*

	a:	**e**:	**i:**	**o:**
Time features	71.1	65.0	66.4	63.5
Frequency features	77.0	75.1	69.0	71.7
All features	72.1	74.0	70.0	65.4

6. References

[1] Atal BS, "Speech analysis and synthesis by linear prediction of the speech wave", The Journal of the Acoustical Society of America. , Vol.50, p.637-655,1971

[2] Ksr Murty B Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition", IEEE signal processing letters. , Vol.13(1), p.52-55,2006.

[3] E. Moore II, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2849–2852, September 2003

[4] K. E. Cummings and M. A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," Journal of the Acoustical Society of America, vol. 98, no. 1, pp. 88–98,1995.

[5] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," IEEE Transactions on Biomedical Engineering, vol. 55, no. 1, pp. 96–107, 2008

[6] Alku, P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech communication, 11(2), 109-118, 1992

[7] J. Fürnkranz, "Round Robin Classification", Applied physics letters.,Vol.2(4), p.721-747, 2000.

[8] Corinna Cortes Vladimir Vapnik, "Support-Vector Networks", Machine learning. , Vol.20(3), p.273-297,1995

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B.Weiss, "A Database of German Emotional Speech," Proc. Ninth European Conf. Speech Comm. and Technology, pp. 1517-1520, 2005.

[10] Kisler, T. and Schiel, F. and Sloetjes, H. (2012): Signal processing via web services: the use case WebMAUS, Proceedings Digital Humanities, Hamburg, Germany, Hamburg, pp. 30-34, 2012

[11] Cassidy, S. and J. Harrington, "Multi-level annotation in the Emu speech database management system", Speech Communication, 33, 61-77, 2001

[12] Bier SD, Watson CI, McCann CM. "Using the perturbation of the contact quotient of the EGG waveform to analyze age differences in adult speech", [J]. J Voice, 2014

[13] A. I. Iliev and M. S. Scordilis, "Spoken emotion recognition using glottal symmetry," EURASIP Journal on Advances in Signal Processing, Article ID 624575, pp. 1–11, 2011

[14] Proakis, J.G. and Manolakis, D.G, "Digital Signal Processing", Upper Saddle River, NJ: Prentice-Hall, pp 910–913,1996

[15] Shapiro, S.S. and Wilk, M.B, "An analysis of variance test for normality (complete samples)", Biometrika, 52, 591–611, 1965.

[16] Bo Chen Guo-Zheng Li Mingyu You, "Multi-class feature selection using Pairwise-class and All-class techniques", 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), p.644-647,2010.