레터논문 (Letter Paper)
방송공학회논문지 제18권 제5호, 2013년 9월 (JBE Vol. 18, No. 5, September 2013)
http://dx.doi.org/10.5909/JBE.2013.18.5.771
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

음성감정인식에서 음색 특성 및 영향 분석

이 정 인[a)‡], 최 정 윤[a)], 강 홍 구[a)]

Analysis of Voice Quality Features and Their Contribution to Emotion Recognition

Jung-In Lee[a)‡], Jeung-Yoon Choi[a)], and Hong-Goo Kang[a)]

요 약

본 연구는 감정상태와 음색특성의 관계를 확인하고, 추가로 cepstral 피쳐와 조합하여 감정인식을 진행하였다. Open quotient, harmonic-to-noise ratio, spectral tilt, spectral sharpness를 포함하는 특징들을 음색검출을 위해 적용하였고, 일반적으로 사용되는 피치와 에너지를 기반한 운율피쳐를 적용하였다. ANOVA분석을 통해 각 특징벡터의 유효성을 살펴보고, sequential forward selection 방법을 적용하여 최종 감정인식 성능을 분석하였다. 결과적으로, 제안된 피쳐들으로부터 성능이 향상되는 것을 확인하였고, 특히 화남과 기쁨에 대하여 에러가 줄어드는 것을 확인하였다. 또한 음색관련 피쳐들이 cepstral 피쳐와 결합할 경우 역시 인식 성능이 향상되었다.

Abstract

This study investigates the relationship between voice quality measurements and emotional states, in addition to conventional prosodic and cepstral features. Open quotient, harmonics-to-noise ratio, spectral tilt, spectral sharpness, and band energy were analyzed as voice quality features, and prosodic features related to fundamental frequency and energy are also examined. ANOVA tests and Sequential Forward Selection are used to evaluate significance and verify performance. Classification experiments show that using the proposed features increases overall accuracy, and in particular, errors between happy and angry decrease. Results also show that adding voice quality features to conventional cepstral features leads to increase in performance.

Keyword : Emotion recognition, voice quality features

I. Introduction

The importance of emotion recognition has become greater recently, as more natural communication between humans and machines becomes desirable. However, defining the task of recognizing emotions from speech is somewhat ambiguous when compared with general speech recognition.

Various studies have shown that measurements related to fundamental frequency (F0) and energy (RMS), often are common prominent features for detecting emotional states

a) 연세대학교 (Yonsei University)
‡ Corresponding Author : 이정인(Lee Jung-In)
E-mail: junginida@dsp.yonsei.ac.kr
Tel: +82-2-2123-4534
· Manuscript received 25, June 2013 Revised 8, August 2013 Accepted 8, August 2013

[1-3]. Features such as MFCCs (Mel-frequency cepstral coefficients) and duration measurements have also been studied [4][5]. Though prosodic features were mainly studied in emotion recognition fields, spectral features also show the efficiency for emotion recognition. The studies on voice quality attempted to find acoustic correlates for different phonation types for emotional speech [2]. The features used in previous studies are insufficient to classify the valence of emotions, though activation of emotion is easily classified. This study focused on the importance of the voice quality to overcome limitations of prosodic features.

This letter aims to expand upon analyses of the relationship between voice quality measurements and emotions. There are three main objectives of this study. First, various voice quality measurements, in addition to prosodic features, are examined for four major emotions. Second, we attempt to quantitatively identify useful measurements for classifying different emotions. The last objective is to apply these measurements to improve emotion classification rate, using various combinations of features.

II. Method

1. Feature measurement

The features examined in this paper can be divided into two groups. The prosodic features group include widely used measurements related to fundamental frequency and energy measurements. The other is the voice quality features group, explained in detail below. Features are extracted over an entire utterance, using a 20msec Hamming window with 50% overlap between adjacent frames. Global feature measurements were selected for utterance-based features, which average out phonemic information, while frame-based cepstral measurements were extracted for comparison. Four global statistics, which are the mean, median (med), standard deviation (std), and inter-quartile range (iqr) of each feature, are obtained. Also, first and second derivative values are found by calculating the absolute difference between the frame preceding and following a current frame.

2. Voice quality features

Open quotient (OQ) is related to the relative duration of glottis opening. It is expected to be greater in a breathy voice, and lesser in pressed voice [6]. In this paper, OQ is defined as the slope between the first and second harmonics (H1 and H2). In order to approximate the glottal source term, harmonics are extracted from the residual signal after 16th-order LPC filtering.

The harmonics-to-noise ratio (HNR) denotes the log ratio of the energies of periodic and aperiodic signal components. By using cepstrum analysis, the convolved excitation and filter are easily separated, and harmonic and noise components are also easily estimated [7].

Spectral tilt describes the amount of decrease in spectral intensity as frequency increases. Two slope measurements are extracted based on previous studies [6].

$$Stilt1 = \frac{H1 - A3}{F3 - F0} \tag{3}$$

$$Stilt2 = \frac{H1 - A2}{F2 - F0} \tag{4}$$

In the equations above, H1 is the amplitude of the fundamental frequency, and A2 and A3 refer to the amplitude of the second and third formants (F2 and F3), respectively.

Spectral sharpness (denoted as Sharp) is computed by equation (5),

$$Sharp(t) = \frac{1}{M}\sum_{n=1}^{M}\left|S(n,t) - S(n-1,t)\right| \tag{5}$$

S(n,t) is spectral amplitude of frequency n at time t, and M is a frequency index corresponds to 2kHz. The harmonic component is more prominence relative to the noise com-

ponent, compared with spectral regions above 2kHz. This measurement expresses the degree of impulse-like characteristics in the glottal source.

Band energy measurements represent the energy in specific frequency bands. Frequency bands are determined by considering typical values of glottal vibration and of formant frequencies. Four frequency bands examined in this study are: 0-250 Hz (E1), 0-500 Hz (E2), 500-1000 Hz (E3), and 2500-3500 Hz (E4). Energy in each band is normalized by overall energy. Since each phonation type exhibits varying patterns of spectral amplitude over different frequency ranges, these measurements are expected to reflect changing voice quality.

III. Feature analysis

1. Database

The speech material used in this study is the Korean Emotion Corpus developed by Kang et al [2]. It includes four emotional states, happy, sad, angry and neutral, recorded by 15 actresses and 15 actors in Korean. 45 dialogue sentences for each emotion were recorded by each speaker at 16kHz. The total database is divided into an analysis and an experiment set. Utterances of five female and five male speakers, totalling 1486 utterances, are used as the analysis set. The rest of the data, including 2969 utterances from 10 male and 10 female speakers.

2. Feature selection

In order to verify the usefulness of voice quality measurements, emotion recognition experiments are conducted next, using the experiment data set. The Sequential Forward Selection (SFS) algorithm is used to select measurements which most contribute to classifying emotion while minimizing correlation among features [8].

Table 1 shows best selected feature sets and resulting recognition rates from the SFS process. First, a basic recognition experiment is implemented using only conventional F0 and energy features. Second, voice quality measurements. The first experiment resulted in a best recognition rate of 49.5% with 5 features. For the prosody and voice quality feature, the recognition rate improves to 62.1% using 14 features. Here, measurements related to voice quality features comprise a larger portion than conventional prosodic feature measurements.

표 1. 시퀀셜 포워드 셀렉션 (SFS)를 이용하여 결정된 피쳐 목록
Table 1 Feature list using sequential forward selection.

step	prosody	voice quality
1	Δ RMS (med)	E2(mean)
2	F0(iqr)	HNR(med)
3	RMS(std)	Δ Δ Stilt2(mean)
4	F0(mean)	Δ E2(mean)
5	Δ RMS(std)	HNR(med)
6	-	OQ(med)
7	-	Δ F0(med)
8	-	Δ Δ HNR(std)
9	-	E1(std)
10	-	Δ OQ(std)
11	-	Δ Δ E2(iqr)
12	-	E2(iqr)
13	-	Δ Δ E4(med)
14	-	E1(iqr)

IV. Experimental results

Performance comparison is carried out using two different classification methods. The utterance-based method extracts features over an utterance to represent long-term characteristics of emotion. The frame-based method uses cepstral features extracted from each frame. Results using the combination of utterance-level features are shown in Table 2. Each feature group is modeled by 8 mixture GMMs for 16 feature vectors. Proposed feature denoted as are selected from the analysis result in previous section.

Proposed features have higher accuracy than conventional prosody features. Relative improvement of overall accuracy is approximately 5%. Furthermore, errors on the emotion valence are decreased, decreased errors between happy and anger are 1.1% and between neutral and sad are 3.1%.

표 3. 16개의 선별된 특징을 이용한 감정인식 결과
Table 3 Confusion matrix of experimental result using 16 prosody and voice quality features. The results of 16 conventional prosody features are marked in brackets.

Emotion	neutral	joy	sad	anger
neutral	72.1 (70.0)	8.0 (3.1)	19.9 (25.6)	0.0 (1.3)
joy	1.3 (8.24)	59.1 (56.9)	13.3 (7.5)	26.3 (27.3)
sad	23.7 (24.2)	10.4 (13.5)	59.7 (57.2)	6.3 (5.12)
anger	2.75 (4.2)	23.6 (24.8)	6.25 (8.9)	67.4 (62.1)
Total	64.58 (61.54)			

표 4. MFCC와 제안된 특성을 조합한 감정인식 성능 비교
Table 4. Results using combination of cepstral features and voice quality features. err(1) and err(2) denote errors of arousal and valence.

Features	MFCC39	MFCC24	MFCC24+ prop.	MFCC39+ prop.
accuracy	74.0	74.2	76.7	76.0
err(1)	20.5	24.2	22.6	25.9
err(2)	16.7	15.8	13.8	12.3

The second experiment uses MFCCs and voice quality features within a frame-based scheme. Since spectral information is contained in MFCCs, spectral tilt and band energies are not used in this experiment. MFCC39 features consist of standard 0th through 12th coefficients with their first and second derivatives. MFCC24 features include the 0th through 7th coefficients with their first and second derivatives. The proposed feature uses F0, OQ, and HNR features with MFCCs.

As shown in Table 4, proposed features improve the accuracy compared to standard MFCC features. These results show that features related to the glottal source provide information complementary to spectral features.

V. Conclusion

This study presents an investigation of the relationship between voice quality and emotions, and useful feature measurements. For voice quality features, open quotient, harmonics-to-noise ratio, spectral tilt, spectral sharpness, and band energy features are considered. From the sequential forward selection algorithm, it is confirmed that voice quality features are able to discriminate emotion valence effectively. Voice quality measurements related to variability were found to contribute significantly as well. Especially, these measurements appeared to be effective for discriminating emotion valence. Detecting emotions on the valence scale is a major difficulty in emotion recognition, so that decreasing total errors in emotion valence leads to overall improvement.

참 고 문 헌

[1] R. Cowie, E. Douglas-Cowei, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, “Emotion Recognition in Human Computer Interaction,” IEEE Signal Processing Magazine, pp. 32-80, 2001.

[2] B.-S. Kang, “Text independent emotion recognition using speech signals,” M. S. Thesis, Yonsei university, 2000.

[3] I. Murray, J. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion,” J. Acoust. Soc. Am, vol. 93 (2), pp. 1097-1108, 1993.

[4] H.-S. Kwak, S.-H. Kim, Y.-K. Kwak, “Emotion recognition using prosodic feature vector and Gaussian mixture model,“ Korean Soc. for Noise and Vibration Eng, pp. 762-765, 2002.

[5] S. Yacoub, S. Simske, X. Lin, J. Burns, ”Recognition of Emotionsin Interactive Voice Response System,” Proceedings of the Eurospeech 2003, Geneva, 2003.

[6] J.-Y. Choi, M. Hasegawa-Johnson, J. Cole, ”Finding intonational boundaries using acoustic cues related to the voice source,” J. Acout. Soc. Am. vol. 118 (4), p. 2579-2587, 2005.

[7] G. de Krom, “A Cepstrum-based technique for determining a Harmonic-to-Noise ratio in speech signals,” J. Speech Hearing Res. vol. 36, pp. 254-266, 1993.

[8] P. Pudil, F. J. Ferri, J. Novovicova, J. Kittler, “Floating Search Methods for Feature Selection with Nonmonotonic Criterion Functions,” Proceedings of the IEEE International Conference on Pattern Recognition, vol. 2, pp. 279-283, Jerusalem, 1994.