논문 2010-1-7

음성을 이용한 감정 정보 추출 방법

An acoustic study of feeling information extracting method

이연수*, 박용범*

Yeon-Soo Lee, Young B. Park

요 약 최근 콜센터 등에서는 고객을 음성 미디어를 통하여 서비스하고 있다. 이런 콜 센터에서 제공하는 다양한 서비스의 품질 측정 방법 중 음성 대화 속 화자의 감정에 따른 측정 방법이 있다. 본 연구에서는 화자의 음성을 이용하여 그 사람의 감정을 알아내고자 하였다. 이를 위하여 음성 신호로부터 여러 가지 파라미터를 추출하고 분석함으로써 인간의 감정을 분류하였다. 사람의 감정은 크게 기쁨, 슬픔, 흥분, 보통 등 4가지 상태로 나눌 수 있다. 대부분의 음성 서비스 품질은 흥분 또는 분노의 상태가 중요하다. 본 논문에서는 이와 같은 감정을 Pitch와 Amplitude를 기초로 한 5가지 요소를 통하여 효율적으로 대화자간의 문제가 되는 대화를 선별해 내는 방법을 연구 하였다.

Abstract Tele-marketing service has been provided through voice media in a several places such as modern call centers. In modern call centers, they are trying to measure their service quality, and one of the measuring method is a extracting speaker's feeling information in their voice. In this study, it is proposed to analyze speaker's voice in order to extract their feeling information. For this purpose, a person's feeling is categorized by analyzing several types of signal parameters in the voice signal. A person's feeling can be categorized in four different states: joy, sorrow, excitement, and normality. In a normal condition, excited or angry state can be major factor of service quality. In this paper, it is proposed to select a conversation with problems by extracting the speaker's feeling information based on pitches and amplitudes of voice.

Key Words : Service quality, Voice pitch, Feeling information.

Ⅰ. 서 론

최근에 화자의 음성을 분석하여 감정의 변화를 인지하는 연구가 증가 하고 있다. 이와 같은 연구가 증가하는 이유는 감정 이라는 부분이 여러 곳에서 사용 될 수 있는 중요한 요소이기 때문이다.[4][1] 일반적으로 감정의 변화를 인지하는 감정인지 시스템은 대화 중 감정 변화가 발생하면 이를 인지하는 방식으로 작동 하여야 한다. 하지만, 슬픔이나 흥분과 같은 감정 상태를 나타내는 음성은 감정인지 시스템에서 구분하기 힘들기 때문에 사람의 개입 없이 신뢰할 수 있는 시스템을 구축하기 어렵다.[6] 따라서 현재 사용되는 감정인지 시스템은 여러 화자간의 대화에서 감정변화를 인지하기 위해서는 감정의 좋고 나쁨을 점수로 나타내는 방식을 사용하고 있다.[2] 또한 시스템의 성능적인 측면을 만족시키기 위해서는 신뢰할 수 있는 감정의 음성적 특징들을 알아야만 한다.[8] 더욱이 감정의 음성적 특징을 이용하면 이미 알고 있는 음성의 특징을 기반으로 하여 본래 음성에 감정적인 요소를 추가하는 변형도 가능하다[9].

본 연구는 콜센터의 고객과 상담원의 대화 중 문제가 되는 대화를 선별하고자 녹음된 대화를 이용하여 실험하였다. 대부분의 콜센터와 같은 상업적 환경에서는 정확한 감정인식보다 최소한의 비용으로 포괄적인 선별이 더

*정회원, 단국대학교 정보아키텍쳐 연구실

접수일자 200x.x.xx, 수정일자 200x.x.xx (기재불가)

욱 중요한 요인이 된다. 따라서 본 연구에서는 어느 정도의 감정인식이 이루어지는 상황에서 최소 비용으로 시스템을 구성할 수 있도록 최소 비용으로 검출 가능한 음성 분석 요소만으로 감정 변화를 인지하고자 하였다.

II. 음성 특징 측정 요소

음성의 특징을 분석하는 방법에는 여러 가지가 있지만 많이 사용되는 음성 정보의 기초 요소가 되는 것으로는 피치(Pitch)와 포만트(Formant) 그리고 소리의 크기를 나타내는 진폭(Amplitude)등이 사용된다. 이들 요소는 보통 자기 상관계수, TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add), DFT (Discrete Fourier Transform)등의 보편적으로 사용되는 일반적인 음성 특징 추출방법을 통해 구한다.[5]

1. 피치(Pitch)

피치는 기본 주파수(Fundamental Frequency)를 의미하며 음성의 주기적 특성을 나타낸다. 하지만 음성은 부분적으로 완전하지 않은 주기성 (Quasi- periodic)을 나타내기 때문에 근사치만을 구할 수 있다. 피치를 구하는 방법으로는 시간 축에서 자기상관계수를 이용하는 방법과 주파수 축에서 DFT를 이용하여 배주파수 성분을 알아내어 구하는 방법이 있다.

2. 포만트(Formant)

인간의 발성기관인 성도의 공명을 나타내는 것으로 음성 신호를 주파수 영역으로 변환하여 주파수 에너지의 정점을 연결한 선들을 말한다. 주로 낮은 주파수 정점부터 F1, F2.. 순으로 표시해 나아간다. 보통 모음이 발생할 때 F1 ~ F3주파수 영역에서 높은 에너지가 나타난다. F1은 턱의 열림과 관계가 있으며, F2는 그림 1의 14 ~ 18 부분의 혀의 위치와 관련이 있다.[10]

3. 진폭(Amplitude)

소리의 크기를 결정하는 요소로 큰 소리는 진폭이 크고 작은 소리는 진폭이 좁아진다. 사람마다 말할 때 소리의 크기에는 많은 편차가 있어 개인적인 차이와 말을 하는 중간에도 소리의 크기는 계속 변화하므로 이에 대한 고려로 평균 소리크기를 이용하기도 한다.[7][3]

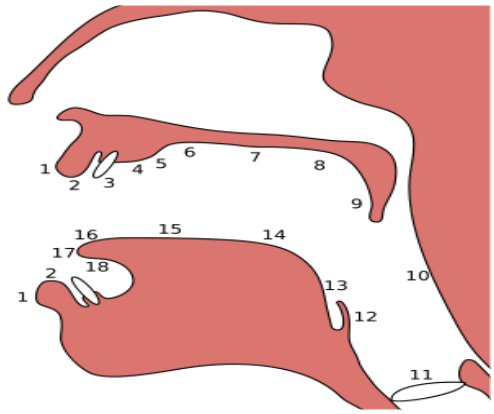

그림 1. 사람의 발성 구조
Fig. 1. Vocal structure of Human

III. 음성 감정 추출 방법

음성에서 감정을 추출하기위하여 본 연구에서 사용한 요소는 다음과 같다;

-. 화자의 발화 속도 변화
-. 화자의 발성 소리 크기의 변화
-. 대화 중 무음 구간의 길이
-. 대화자 간의 중복 발성 구간의 길이
-. 대화자 간의 발성 비율

화자의 발화 속도 변화를 알아내기 위해서 유성음(Voiced sound)의 발생 속도를 측정하여야 한다. 일반적으로 유성음은 주기성을 가지며 비교적 뚜렷한 피치를 보이므로 이러한 성질을 기반으로 발화 속도를 측정하였다.

화자의 발성 크기는 소리의 진폭이나 에너지를 이용하여 측정이 가능하다. 대화 중 무음 구간과 대화자 간의 중복 발성 구간의 길이 또한 소리의 진폭이나 에너지를 분석함으로서 측정가능하다. 대화자간의 발성 비율은 전체 대화가 끝난 후 각 화자의 발성 시간의 비율로 구해지므로 별도의 분석 없이 얻을 수 있다.

따라서 음성 감정 추출을 위하여 피치와 소리의 진폭 또는 에너지를 구해야 하는데, 여기에는 DFT와 자기 상관계수가 이용된다. DFT는 다음과 같은 식 1에 의해 구해진다. 역 DFT는 식 2와 같다.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} k_n}$$
$$where \ \ k = 0, 1, \cdots, N-1 \qquad (1)$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{2\pi i}{N} k_n}$$
$$where \ \ n = 0, 1, \cdots, N-1 \qquad (2)$$

자기상관계수는 식 3) 의해 구해진다.

$$Rxx(j) = \sum_{n} x_n \overline{x}_{n-j} \qquad (3)$$

소리의 진폭 또는 에너지는 DFT의 X_0를 이용하거나 자기상관 계수 $R_{xx}(0)$를 이용한다. 하지만 DFT를 이용하거나 자기상관계수를 이용하는 방법은 계산 량이 많아 콜센터와 같이 컴퓨팅 파워를 여러 응용프로그램과 나누어 써야 하는 환경에서는 사용이 어렵다. 본 연구에서는 유성음의 신호적 특성이 무성음에 비해 상대적으로 저주파 영역에 놓이고 높은 에너지를 가짐에 착안하여 계산량이 적은 Zero-crossing을 이용하였다. Zero-crossing은 식 4와 같이 구할 수 있다. 또한 소리의 진폭도 식 5와 같이 구하여 진폭만을 고려한 방법을 도입하였다.

$$Zcr = \frac{1}{N-1} \sum_{n=1}^{N-1} L(x_n \cdot x_{n-1})$$
$$where \ \ L(x) = \begin{cases} 1 : x < 0 \\ 0 \ otherwise \end{cases} \qquad (4)$$

$$Amp = x_{\max} - x_{\min} \qquad (5)$$

본 연구에서 사용된 음성 자료는 실제 콜센터에서 녹음된 H사와 N사의 고객과 상담원 사이에서 녹음한 전화 음성을 사용 하였다. 전화 음성은 8KHz sampling rate, 16bit, Stereo(상담원-좌측, 고객-우측) 형태로 녹음되었다. 이 음성 자료를 처리하기 위하여 그림 2에서처럼 두 채널을 나누고 각 채널을 음성 특질이 유지되는 시간 세그먼트(Segment)로 분할하여 처리하였다. 각 시간 세그먼트 별로 신호처리를 한 결과 유사 특질이 유지되는 세그먼트들을 묶어 하나의 음성 블록(Speech Block)으로 합쳐 음성 블록 단위로 사용하였다.

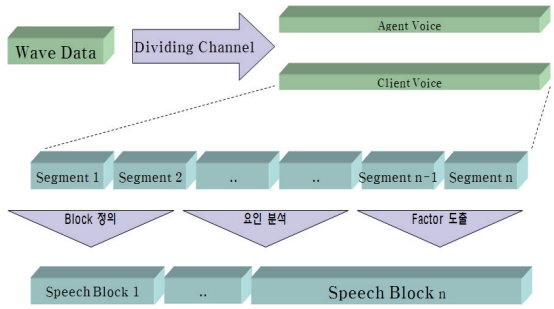

그림 2. 음성 세그먼트와 음성 블록
Fig. 2. Voice Segments and Voice Blocks

그림 3에서처럼 음성 블록 중 안정적인 주기성을 보이며 음성 에너지가 높은 음성 블록을 유성음 음성 블록으로 간주하였고 이를 기반으로 발성 속도를 측정 하였다. 발성속도는 시간당 유성음 음성 블록 수로 정의 된다.

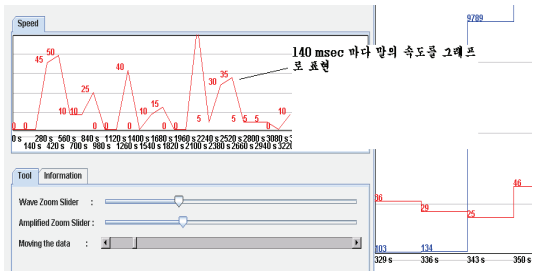

그림 3. 발성 속도의 측정
Fig. 3. Measuring Voice speed

소리크기의 변화를 측정하기 위해 각 음성 세그먼트 단위로 식 1의 DFT를 수행한 후 X_0를 이용하는 방식과 식 5와 같이 음성 세그먼트 내 최대값과 최소값을 찾아 그 차이를 이용하였다. 각 화자별 평균 소리크기를 알기 위해 대화 초기의 발성이 이루어지는 30개 음성 세그먼트의 평균을 구하고 식 6의 이동 평균이용 하였다.

$$R\_Avg_t = \frac{N \times R\_Avg_{t-1} + x_t - x_{t-N}}{N} \qquad (6)$$

여기서 N은 이동 윈도우 크기이다. 이 이동 평균의 크기를 이용하여 화자의 발성 소리 크기의 변화, 대화 중 무음 구간의 길이, 대화자 간의 중복 발성 구간의 길이를

구하였다.

Ⅳ. 실험과 결과

앞에서 논의한 방식으로 음성 감정을 추출하였고 추출을 위하여 일반적으로 많이 사용되는 DFT와 식 4와 식 5에 제시한 Zero-crossing과 소리 진폭 측정 방식을 이용하여 컴퓨팅 파워를 적게 사용하는 제안 방식을 비교 실험하였다. 그림 4는 구현된 실험 도구이다.

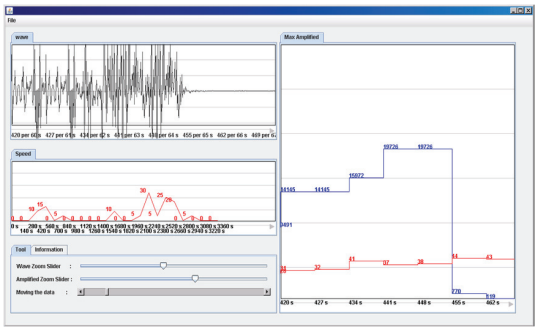

그림 4. 음성 감정 정보 추출 도구
Fig. 4. Voice feeling extraction tool

실험을 위하여 평균 5분 정도의 선택된 대화를 사용하였다. 이 중 약 1/3의 문제가 있는, 즉 감정이 격하게 변하는 대화를 미리 선발하여 포함 시켰다. 각 대화는 문제의 심각 정도에 따라 1~10점까지 등급을 주었고 일정 점수 이상을 문제가 있는 대화로 취급하였다.

DFT를 이용한 음성 감정 정보 추출과 제안한 식 4와 식 5의 Zero-crossing과 소리 진폭 측정 방식을 적용한 방식의 음성 감정 정보 추출을 통한 인식 율은 <표 1>과 같았다. 여기서 "정 인식 율"은 문제가 있는 대화를 얼마나 검출해내었는가를 의미하며 "오 인식 율"은 문제가 없는 대화를 문제 대화로 잘못 인식 했는가를 나타낸다. 즉 "정 인식 율"은 문제가 있는 대화이었음에도 불구하고 정상대화로 인식하는 가에 대한 척도이며 "오 인식 율"은 정상 대화를 문제가 있는 대화로 잘못 인식하는 정도를 나타낸다. DFT를 이용한 방법이 제안된 방식보다 정 인식 율이 다소 높은 것은 주파수 영역에서 음성 정보를 많은 컴퓨팅 자원을 이용하여 정상적으로 처리한 것이 최소한의 컴퓨팅 자원을 사용하는 제안한 방식보다 정밀한 비교가 가능했었기 때문이라고 사료 된다.

표 1. 음성 감정 정보 인식 율 비교
Table 1. Comparison of Voice feeling information

	DFT	제안된 방식
정 인식 율	76%	74%
오 인식 율	3%	3%

"정 인식 율"을 높이기 위해 문제대화 인식 등급을 조정할 수 도 있지만 이는 "오 인식 율"의 증가로 이어진다. "정 인식 율"을 다소 포기하고 "오 인식 율"울 낮춘 이유는 본 연구의 목적인 콜 센터의 요구사항이 정상대화를 문제대화로 인식하는 오류를 최소화 하는 것이었기 때문이다.

문제 대화 중 가장 선별하기 어려웠던 것은 감정의 변화가 발생하여도 즉 감정이 격해지거나 화를 내는 상태가 되어도 오히려 차분하고 냉정한 어투를 사용하는 경우였다.

실험 결과 DFT와 제안된 방식으로 음성 감정 정보를 추출한 인식 율이 거의 유사하여 콜센터와 같은 저 컴퓨팅 사양에서는 제안된 방식을 이용하여도 큰 인식 율의 손실 없이 문제 대화를 추출 할 수 있었다.

Ⅴ. 결 론

화자의 음성을 분석하여 감정변화를 인지하고 이를 기반으로 대화를 분석하여 문제 대화를 추출하는 방법에 관하여 연구하여 보았다. 언어적 정보와 비언어적 정보 모두 이용이 가능하지만 아직 한국어 음성인식이 완전하게 이루어지지 않는 상황에서 그 내용적 의미를 파악하여 감정 변화를 인지하는 것은 매우 어려운 문제이며 많은 컴퓨팅 파워를 요구하기 때문에 상업적 응용도 쉽지 않은 상황이다.

본 연구에서는 적은 컴퓨팅 파워를 사용하여 받아들일만한 수준의 인식 율을 보이는 방법을 찾아보았다. 물론 한국어 음성을 기준으로 한 연구이지만 언어 종속적인 부분이 적어 다른 언어에도 쉽게 적용이 가능할 것으로 보인다. 현실적인 영업상 보안 문제로 많은 대화 샘플 데이터를 구할 수 없어 충분한 실험을 하는데 제한이 있었다. 앞으로 조금 더 많은 컴퓨팅 자원을 이용할 수 있다면 인식을 위한 단순한 어댑티브 알고리즘을 적용해

상황 정보를 추가해 보는 것이 다음 연구가 될 것이다.

참 고 문 헌

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, and Taylor, J., "Emotion recognition in human-computer interactioins", IEEE Sig. Proc. Mag., vol.18(1), pp. 32-80, Jan 2001

[2] Davis, C., Kim, J., Grauwinkel, K. and Mixdorff, H., "Lombard speech: Auditory(A), Visual(V) and AV effects", Proceedings of Speech prosody, pp.361-365, Dresden, Germany, 2006.

[3] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Classification of Speech under Stress based on Features derived from the Nonlinear Teager Energy Operator," Proc. of ICASSP, pp. 549-552, 1998.

[4] Lee, C. M., and Narayanan, S., "Towards detecting emotion in spoken dialogs," IEEE Trans. on Speech and Audio Processing, Vol. 13(2), pp.293-303, 2005

[5] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993

[6] Patil, V. and Rao. P., "Acoustic cues to manner of articulation of obstruents in Marathi", Proc. of frontiers of research on Speech & Music, Kolkata, India, February 2008.

[7] S. Fukuda and V. Kostov, "Extracting Emotion from Voice," Proc. of IEEE, pp. IV-299-304, 1999.

[8] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," Proc. of Internet Conf. on Spoken Lang. Processing. (ICSLP 04), Vol. 1, pp. 2193-2196, Jeju Island, Korea, 2004,

[9] Cahn, J.E., "Generating Expressions in Synthesized Speech",Master's Thesis MIT, 1989.

[10] 양병곤, 프라트를 이용한 음성분석의 이론과 실제, 만수출판사, 2003

※ 본 연구는 2008년도 단국대학교 대학연구비의 지원으로 연구되었음

저자 소개

이 연 수

- 단국대학교 공학대학 컴퓨터과학과 학사
- 단국대학교 전자계산학과 컴퓨터과학 석사 재학중

<주관심분야 : 인공지능, 웹 서비스, 시멘틱 웹>

박 용 범(정회원)

- 서강대학교 전자계산학과 학사
- Polytechnic University Computer Science M.S.
- Polytechnic University Computer Science Ph.D.
- (현) 단국대학교 컴퓨터과학 교수

<주관심분야 : 지능형 SE, 프로젝트 관리, 크라우딩 컴퓨팅>