한국어 문미억양 강조를 통한 향상된 음성문장 감정인식

이태승°, 박미경, 김태수
한국과학기술연구원
thestaff@hitel.net, miky@kist.re.kr, ktaesoo@kist.re.kr

Toward More Reliable Emotion Recognition of Vocal Sentences by Emphasizing Information of Korean Ending Boundary Tones

Tae-Seung Lee°, Mikyong Park, Tae-Soo Kim
Korea Institute of Science and Technology

요약

인간을 상대하는 자율장치는 고객의 자발적인 협조를 얻기 위해 암시적인 신호에 포함된 감정과 태도를 인지할 수 있어야 한다. 인간에게 음성은 가장 쉽고 자연스럽게 정보를 교환할 수 있는 수단이다. 지금까지 감정과 태도를 이해할 수 있는 자동시스템은 발성문장의 피치와 에너지에 기반한 특징을 활용하였다. 이와 같은 기존의 감정인식 시스템의 성능은 문장의 특정한 억양구간이 감정과 태도와 관련을 갖는다는 언어학적 지식의 활용으로 보다 높은 향상이 가능하다. 본 논문에서는 한국어 문미억양에 대한 언어학적 지식을 피치기반 특징과 다층신경망을 활용하여 구현한 자동시스템에 적용하여 감정인식률을 향상시킨다. 한국어 감정음성 데이터베이스를 대상으로 실험을 실시한 결과 4%의 인식률 향상을 확인하였다.

I. 서론

인간은 두 가지 방식을 통해 의사를 전달한다. 하나는 특정한 목표를 지닌 정보에 대한 표면적인 신호를 전달하는 것이고, 다른 하나는 그 자신에 대한 암시적인 신호를 전달하는 것이다[1]. 표면적인 신호가 실질적인 행동의 실시를 요구하는 데 비해 암시적인 신호는 표현자의 욕구동기를 이해해 주길 바란다. 지금까지 수행된 대부분의 연구가 전자의 표면적인 신호를 분석하여 문법적 규칙을 알아내는 데 집중되어 왔다. 그러나 사회적 관계라는 측면에서 볼 때 타인의 욕구동기를 이해하지 못하고서는 자발적 협력을 획득할 수 없다. 유비쿼터스 세계의 도래와 함께 인간과 호흡하는 자율장치는 암시적인 신호, 즉 감정과 태도를 지각하는 능력을 구비해야 한다.

음성은 이 두 가지 신호를 교환하는 가장 간단하고 자연스러운 수단이다. 지난 30년 간 인간이 발성한 표면적인 신호를 분석하여 언어적 구조를 밝히기 위해 자동 음성인식 시스템에서 많은 연구가 수행되었다[2]. 이해 비해 구어의 준언어적 의미처리를 위해 암시적 신호에 관심을 기울이기 시작한 것은 불과 몇 년 되지 않는다[3-4]. 준언어적 정보의 처리를 위한 대표적인 도구는 발성된 문장의 피치와 에너지이다.

준언어적 정보의 측정은 정적인 방법과 동적인 방법으로 구분할 수 있다. 전체 문장에 대해 정적인 방법에서 감정과 태도를 인식하기 위한 특징은 피치와 에너지의 통계치이고[5] 동적인 방법에서는 피치 및 에너지의 몇 개 미분차원으로 형성되는 음조곡선이다[6]. 최근 들어 언어적 정보를 감정인식에 활용하려는 시도가 있었고 정적 및 동적 방법에 비해 높은 인식률을 달성한 것으로 보고되었다[7].

위 연구결과[7]는 감정과 태도를 인식할 때 신뢰성을 향상시키기 위해 언어적 지식을 준언어적 정보처리에 적용할 필요성을 제기한다. O Connor와 Arnold[8]에 따르면 영어에는 일곱 개의 핵억양이 존재하고 각 억양은 화자의 진지함, 냉정함, 놀람 같은 감정적 태도적 상태를 나타낸다. 전선아[9]와 이호영[10]은 한국어 문장에서 문미억양의 의미가 감정 및 태도와 특별한 관계가 있다고 주장했다. 이들의 연구결과에 입각할 때 영어의 핵억양과 한국어의 문미억양에 더 높은 중요도를 부여한다면 감정과 태도의 더욱 신뢰할만한 인식이 가능할 것이다.

본 논문에서는 한국어 문장의 문미억양에서 추출한 통계치에 더 높은 가중치를 부여하여 정적인 방법에 기반한 기존의 자동감정인식 시스템의 성능을 향상시킨다. 이 감정인식 시스템에서 각 한국어 문장은 앞 부분의 본체와 뒤 부분의 미부로 분할된다. 본체는 미부의 두 배 길이를 가지며 미부는 문미억양을 포함한다. 한국어 감정음성 데이터베이스에 대해 최고의 감정인식률을 획득하기 위해 본체와 미부에 부여되는 가중치의 최적비율을 탐색한다. 기존의 감정인식 시스템에 이 최적비율을 적용할 때 인식률의 향상을 확인한다.

II. 감정에 대한 한국어 문미억양의 영향

인간은 음성, 표정, 눈물, 몸짓 같은 수단을 통해 자신의 감정을 표현하는데, 이들 가운데 음성은 가장 자연스러우면서 융통성이 풍부하다. 어떤 감정의 내부 원인과 외부 표현이 언제나 합치되는 것은 아니지만 청자는 내재된 감정이 어떤 것인지 청취한 음성을 통해 추측하려고 시도한다. 억양은 감정상태를 음성으로 표현하는 데 있어 중요한 수단이다. 이 절에서는 영어와 한국어에 대해 문헌에서 보고된 억양과 감정 사이의 관계를 살펴본다.

억양과 이에 상응하는 의도 사이의 관계는 억양음운론에 대한 관심이 고조되면서 연구되기 시작하였다. O'Connor와 Arnold[8]는 영어문장이 하나 이상의 억양으로 구성되며 각 억양은 앞머리억양, 머리억양, 핵억양으로 이루어진다고 말했다. 영어에서 핵억양의 종류로는 낮내림조, 높내림조, 낮오름조, 온오름조, 노르내림조, 내리오름조, 가운데수평조의 일곱 가지가 존재한다. O'Connor와 Arnold는 핵억양이 진지함, 냉정함, 놀람과 같은 화자의 감정 및 태도와 관련이 있다고 주장했다. 이들의 관점에 따르면 감정 및 태도에 대해 억양이 밀접한 관련을 맺고 있음을 추론할 수 있다.

Perrehumbert와 Hirschberg[10]가 주도한 또 다른 관점의 억양음운론에서 문장은 하나 이상의 억양구를 포함할 수 있다. 경계억양은 억양구의 마지막 음절에서 실현되고 표시된다. 경계억양에는 총 아홉 가지 종류가 있으며 이들을 두 집단으로 구분할 수 있다. 첫 번째 집단의 경계억양은 L%, HL%, LHL%, HLHL%, LHLHL%의 꼬리표로 표시하고 두 번째 집단은 H%, LH%, HLH%, LHLH%로 표시한다. 여기서, 기호 L은 낮은 억양, H는 높은 억양을 나타내고 %는 문장의 끝을 의미한다. 주목해야 할 것은 경계억양이 핵억양과 동일한 개념이라는 것이다. Pierrehumbert와 Hirschberg는 이들 경계억양이 서술문, 명령문, 의문문 같은 문장의 종류뿐 아니라 감정, 태도, 양상과 같은 화용론적 의미

에 대한 정보까지 전달한다고 주장하였다.

한국어의 특성 중 하나는 문법적으로 가장 중요한 요소가 다른 요소의 오른쪽에 오는 왼쪽 가지치기이다. 이러한 특성으로 문장의 마지막에 위치한 경계억양은 전체 문장의 의미와 결속될 가능성을 갖는다. 따라서 한국어에서 문미억양은 억양과 이의 의미 사이의 관계를 이해하는 데 중요한 역할을 맡는다.

지난 몇 년 간 일단의 한국어 연구자가 화자의 감정과 태도에 문미억양이 미치는 영향을 조사하였다. 전선아[9]는 몇 가지 한국어 문미억양의 의미를 서술하였는데, 그 중에서 감정이나 태도와 연관된 문미억양의 패턴을 다음과 같이 정리하였다.

1) LHL% - 설득, 주장, 확신, 귀찮음, 자극
2) LH% - 귀찮음, 불쾌, 불신
3) HLH% - 자신에 차 있어 청자의 동의를 구함
4) LHLH% - 귀찮음, 자극, 불신
5) HLHL% - 확신, 주장, 꺼림, 설득
6) LHLHL% - 귀찮음 보다는 강한 감정

이호영[11]은 억양음운론을 한국어에 적용하여 문미억양의 태도적 기능을 연구하였다. 이호영에 따르면 한국어 문미억양의 기능은 감정과 태도와 결합하여 다음과 같이 발현된다.

1) 낮은수평조 - 단호, 냉정
2) 높은수평조 - 흥미, 놀람

지금까지 살펴본 억양음운론 관련 연구결과에서 경계억양은 특정 화자가 발성한 문장의 감정을 파악하는 데 매우 중요한 역할을 한다는 것을 알 수 있다. 특히 한국어에서 억양에 의한 영향은 문미억양에 집중된다. 그러므로 영어의 경계억양과 한국어의 문미억양에 보다 큰 중요성을 부여한다면 보다 정확하게 화자의 감정을 인지할 수 있을 것이다. 다음 절에서는 한국어 문미억양에 대한 억양음운론적 지식을 도입하여 감정인식능력의 향상에 문미억양이 미치는 영향을 실험을 통해 확인한다.

III. 실험

2절에서 살펴본 언어학적 지식을 자동감정인식 시스템에 접목하기 위해 기존의 정적특징에 기반한 시스템을 구현하고, 이 시스템이 문미억양을 포함하는 음성구간에서 계산된 통계치에 다른 구간의 통계치보다 높은 중요성을 두도록 했다. 이 시스템의 성능을 한국어 감정음성 데이터베이스에 대해 평가하고 문미억양에 가해지는 최적의 가중치를 탐색하여 최고의 인식률을 달성하도록 한다.

3.1. 구현시스템

본 논문에서 구현된 감정인식 시스템은 감정 파라미터로 피치관련 특징을, 패턴인식 방법으로 다층신경망을 사용한다. 이 시스템에서는 음성문장 검출, 음성문장에서 피치곡선 추출, 감정특징 계산, 감정 학습 및 인식의 네 단계를 거친다.

감정을 인식하기 위한 특징이 음성문장에 대해 계산되어야 하기 때문에 제일 먼저 입력된 소리신호에서 문장을 검출해야 한다. 이를 위해 Rabiner와 Sambur의 방법을 바탕으로 한 끝점검출 알고리즘을 구현시스템에 도입하였다. 이 알고리즘은 음성검출 파라미터로 신호의 에너지와 영교차율을 사용하고 실시간 처리 특성을 갖는다. 이 알고리즘에서 거치는 처리상태는 총 일곱이며 그림 1에서 각 상태와 상태간 천이를 보여준다. 이 그림에서 START은 알고리즘의 시작, SILENCE는 묵음처리, VL_START_SUS는 무성음개시 의심, V_START_SUS는 유성음개시 의심, S_START_DET는 음성개시 확정, S_END_SUS는 음성종료 의심, STOP은 음성종료 확정을 가리킨다.

피치곡선은 이전 단계에서 검출된 문장에서 추출된다. 구현시스템의 피치추출 알고리즘은 David와 Niederjohn이 제안한 알고리즘[13]을 일부 수정하여 도입한 것이다. 원형 피치추출 알고리즘은 연속적인 30 ms 길이의 음성구간에 대해 표준 단구간 자기상관함수를 계산하고 각 음성구간은 전체 길이의 66.7%가 겹치도록 구성되었다. 이 원형 알고리즘을 저주파수대역과 고주파수대역의 에너지 스펙트럼을 활용하여 신호의 유성음 여부를 결정하도록 수정하였다. 고대역의 에너지가 저대역보다 크다면 해당 음성구간은 유성음이 아닌 것으로 판정할 수 있으므로 피치추출 처리를 생략한다.

정적방법에 기반한 많은 감정인식 시스템에서 피치와 더불어 음성에너지를 감정특징의 원소로 사용한다. 그러나 음성에너지는 마이크 증폭이나 음원과 청원 간 거리의 영향으로 왜곡되기 쉽기 때문에 본 시스템에서는 피치관련 특징만 사용한다. 구현시스템에 사용한 열 두 개의 피치파생 특징은 [5]와 [7]에서 사용된 특징 중에서 선택하였다. 이들을 표 1에서 설명한다.

다층신경망은 위의 열 두 특징을 입력패턴으로 하여 화자의 전형적인 감정을 학습하고 새로 입력되는 패턴을 학습된 감정으로 분류한다. 다층신경망은 하나의 입력계층, 0개 이상의 은닉계층, 하나의 출력계층을 갖는다[14]. 입력계층은 패턴을 받아들이며, 은닉계층은 신경망 동작을 위한 학습능력을 결정하고, 출력계층은 대상감정의 인식점수를 산출한다. 각 계층은 하나 이상의 계산노드로 구성되며 한 계층 내의 모든 노드는 인접계층의 노드와 완전하게 연결된다. 다층신경망의 학습에 보편적으로 사용되는 오류역전파 알고리즘은 최대기울기감소 방법에 근간을 둔다[14]. 오류역전파 알고리즘은 다층신경망에게 기대된 동작을 실현하여 입력패턴을 대응감정으로 분류한다. 이는 다층신경망의 현재출력과 희망출력 간의 오차를 출력계층에서 입력계층으로 역전파시키면서 이 오차가 최소가 되도록 다층신경망의 내부연결 가중치들을 조절하는 방법으로 실현된다.

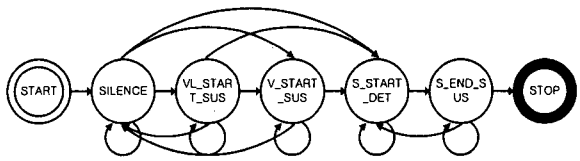

그림 1. 입력음성신호에서 음성문장 검출을 위한 상태도

표 1. 구현시스템에서 사용된 피치파생 특징

특징번호	설명
1	평균피치
2	피치 표준편차
3	최대피치
4	전체문장 내에서 최대피치의 상대적 위치
5	최소피치
6	전체문장 내에서 최소피치의 상대적 위치
7	최대상승기울기
8	평균상승기울기
9	최대하강기울기
10	평균하강기울기
11	평균기울기
12	밀변이 최소피치 위치인 피치곡선으로 형성된 면적

3.2. 감정음성 데이터베이스

본 논문에서 사용된 감정음성 데이터베이스는 화자 및 문장 독립적인 감정인식 시스템의 개발과 시험을 위해 설계되었다. 이 데이터베이스가 목표로 한 감정은 기쁨, 슬픔, 분노, 중성의 네 감정이다. 문장은 45개의 단순 한국어 서술문이고 각 문장을 한국인 아마추어 배우인 세 명의 남성과 세 명의 여성이 세 번씩 발성하였다. 발성된 각 문장은 16 kHz 샘플링 속도와 16 bit 양자화 크기로 설정된 디지털 오디오 테이프레코더를 사용하여 조용한 사무실 환경에서 녹음되었다. 각 문장 앞뒤로 약 50 ms 길이의 묵음이 포함된다. 이 데이터베이스에 대해 또 다른 서른 명의 한국인 청취자에게 행해진 청취실험에서 78.2%의 인식률이 기록되었다.

3.3. 실험조건

문미억양이 감정인식에 미치는 영향을 확인하기 위해 세 가지 실험조건을 설정하였다. 첫째, 감정음성 데이터베이스에 대한 기준성능을 얻기 위해 일반평가가 수행된다. 이 조건에서 인식되는 감정은 다음의

식으로 결정된다.

$$L = SelMax\left(M_i\left(F_{Whole}\right)\right), \quad i \in \{기쁨, 슬픔, 분노, 중성\} \quad (1)$$

여기서 F_{Whole} 은 전체 문장의 특징벡터를, M_i 는 대상감정에 대한 출력벡터를, $SelMax$는 최대출력을 갖는 감정 L 을 선택하는 함수를 나타낸다. 둘째, 문장을 앞 부분의 본체와 뒤 부분의 미부로 나누되 본체가 미부의 두 배 길이를 갖도록 하여 문미억양을 포함하는 미부가 본체보다 더 높은 중요성을 갖는지 확인한다. 이 조건에서 인식되는 감정은 다음의 식으로 결정된다.

$$\begin{cases} L_{Body} = SelMax\left(M_i\left(F_{Body}\right)\right) \\ L_{Tail} = SelMax\left(M_i\left(F_{Tail}\right)\right) \end{cases}, \quad i \in \{기쁨, 슬픔, 분노, 중성\} \quad (2)$$

여기서 L_{Body} 는 본체의 특징벡터 F_{Body} 에 대해 선택되는 감정이고, L_{Tail} 은 미부의 특징벡터 F_{Tail} 에 대해 선택되는 감정이다. 셋째, 최고의 감정인식율을 얻기 위해 본체와 미부의 인식결과에 부여되는 최적의 가중치 비율을 탐색한다. 이 조건에서 인식되는 감정은 다음의 식으로 결정된다.

$$L = SelMax\left(a \cdot M_i\left(F_{Body}\right) + b \cdot M_i\left(F_{Tail}\right)\right), \quad i \in \{기쁨, 슬픔, 분노, 중성\} \quad (3)$$

여기서 a 와 b 는 각각 본체와 미부에 대한 다층신경망의 출력벡터에 부여되는 가중치이다.

결과는 감정음성 데이터베이스 내 45개 문장을 모두 사용하여 화자마다 평가된다. 각 문장과 감정에 대한 세 번의 발성 중 둘은 다층신경망의 학습에 사용되고 나머지 하나는 시험에 사용된다. 결과적으로 학습에는 360개(45문장 * 2회 * 4감정) 문장이 사용되고 시험에는 180개(45 * 1 * 4) 문장이 사용된다. 제시된 모든 결과는 다층신경망의 10번에 걸친 학습과 시험으로 얻어진 결과의 평균치이다.

이 실험에 사용된 다층신경망은 입력계층, 하나의 은닉계층, 출력계층으로 구성된다. 입력계층은 피치파생 특징과 대응하는 12개의 입력지점을 갖고, 은닉계층은 30개의 노드를 포함하며, 출력계층은 인식할 감정과 대응하는 4개의 노드를 포함한다. 학습과 시험에 사용되는 모든 패턴은 -1.0과 +1.0 사이의 값으로 평준화된다. 오류역전파 알고리즘의 학습 파라미터로는 0.05의 학습률과 0.05의 목표오류에너지가 지정된다.

3.4. 결과

상기한 세 가지 조건으로 얻은 실험결과를 그림 2에서 정리한다. 이 그림에서 첫 번째 줄은 일반평가에 대한 결과이고, 두 번째와 세 번째는 음성문장의 분리실험에서 본체와 미부 대한 것이며, 네 번째는 본체와 미부에 대해 탐색한 최적의 가중치 비율에 대한 것이다. 여기서 최적 가중치 조건으로 얻은 전체 인식률이 일반조건으로 얻은 인식률보다 4% 높을 뿐 아니라 모든 감정에 대해 인식률이 향상되었음을 알 수 있다.

다양한 가중치 조합을 본체와 미부에 적용했을 때 인식률의 변화를 그림 3에 표시했다. 이 그림에서 보다 높은 중요도가 미부에 가해졌을 때 인식률이 일반조건일 때보다 높아지고 [0.5, 0.5]의 가중치 조합에서 시스템이 가장 높은 인식률을 달성하는 것을 알 수 있다.

	중성	기쁨	슬픔	분노	전체
□일반조건	75.7%	58.3%	67.4%	67.5%	67.2%
▨본체만	71.0%	55.9%	55.7%	60.3%	60.7%
▩미부만	75.7%	52.4%	69.4%	73.8%	67.8%
■가중치 조합	81.9%	58.4%	71.6%	73.0%	71.2%

그림 2. 세 가지 실험조건에 의한 감정 별 평균 인식률과 전체 감정에 대한 평균 인식률

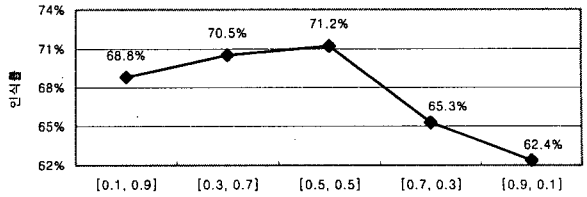

그림 3. [본체, 미부] 형식의 다양한 가중치 조합에 대한 인식률 추이

IV. 결론

지금까지 한국어 문미억양에 대한 억양음운론적 지식을 정적특징에 기반한 자동감정인식 시스템에 적용하여 감정인식의 신뢰성을 향상시키기 위한 방법을 연구하였다. 한국어 문미억양이 감정적 행동에 미치는 영향은 문미억양을 포함하는 미부의 더 높은 인식률을 통해 증명되었다. 이러한 결과를 바탕으로 미부에 더 높은 중요도를 부여했을 때 일반조건일 때보다 4%의 인식률 향상을 얻을 수 있었다. 이와 같은 결과는 억양음운론을 실용적 감정인식 시스템에 도입하는 것에 대해 상당한 의의와 필요성을 제기한다. 향후 연구에서는 자동감정인식에서 보편적인 경계억양의 역할을 한국어뿐 아니라 영어의 음성문장에 대해서도 보다 심도 깊게 연구할 필요가 있다.

참고문헌

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J.G., "Emotion Recognition in Human-Computer Interaction," IEEE Signal Processing Magazine, Vol. 18, No. 1, pp. 32 - 80, Jan 2001.

[2] Gauvain, J. and Lamel, L., "Large-Vocabulary Continuous Speech Recognition: Advances and Applications," Proceedings of the IEEE, Vol. 88, No. 8, pp. 1181 - 1200, Aug 2000.

[3] Yoshimura, T., Hayamizu, S., Ohmura, H., and Tanaka, K., "Pitch Pattern Clustering of User Utterances in Human-Machine Dialogue," Proceedings of the International Conference on Spoken Language, Vol. 2, pp. 837 - 840, Oct 1996.

[4] Dellaert, F., Polzin, T., and Waibel, A., "Recognizing Emotion in Speech," Proceedings of the International Conference on Spoken Language, Vol. 3, pp. 1970 - 1973, Oct 1996.

[5] Bhatti, M. W., Wang Y., and Guan. L., "A Neural Network Approach for Human Emotion Recognition in Speech," Proceedings of the 2004 International Symposium on Circuits and Systems, Vol. 2, pp. 181 - 184, May 2004.

[6] Schuller, B., Rigoll, G. and Lang, M., "Hidden Markov Model-Based Speech Emotion Recognition," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 1 - 4, Apr 2003.

[7] Schuller, B., Rigoll, G. and Lang, M., "Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 577 - 580, May 2004.

[8] O'Connor, J. D. and Arnold G. F., Intonation of Colloquial English, Longmans, 1961.

[9] Jun, S., K-ToBI Labelling Conventions, Ver. 3.1, http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html, 2000.

[10] Pierrehumbert, J. and Hirschberg, J., "The Meaning of Intonation Contours in the Interpretation of Discourse," Intentions in Communication, MIT Press, pp. 271-323, 1990.

[11] 이호영, "한국어의 억양체계," 언어학, 제13호, pp. 129-151, 12월 1991년.

[12] Rabiner, L. and Sambur, M. "An Algorithm for Determining the Endpoints of Isolated Utterances," Bell System Technical Journal, Vol. 54, pp. 297-315, Feb 1975.

[13] Krubsack, D. A. and Niederjohn, R. J., "An Autocorrelation Pitch Detector and Voicing Decision with Confidence Measures Developed for Noise-Corrupted Speech," IEEE Transactions on Signal Processing, Vol. 39, No. 2, Feb 1991.

[14] Bengio, Y., Neural Networks for Speech and Sequence Recognition, International Thomson Computer Press, 1995.