말소리와 음성과학 제1권 제2호 (2009)
pp. 15~22

영어 강세 교정을 위한 주변 음 특징 차를 고려한 강조점 검출

Prominence Detection Using Feature Differences of Neighboring Syllables for English Speech Clinics

심 성 건[1] · 유 기 선[2] · 성 원 용[3]
Shim, Sunggeon · You, Kisun · Sung, Wonyong

ABSTRACT

Prominence of speech, which is often called 'accent,' affects the fluency of speaking American English greatly. In this paper, we present an accurate prominence detection method that can be utilized in computer-aided language learning (CALL) systems. We employed pitch movement, overall syllable energy, 300-2200 Hz band energy, syllable duration, and spectral and temporal correlation as features to model the prominence of speech. After the features for vowel syllables of speech were extracted, prominent syllables were classified by SVM (Support Vector Machine). To further improve accuracy, the differences in characteristics of neighboring syllables were added as additional features. We also applied a speech recognizer to extract more precise syllable boundaries. The performance of our prominence detector was measured based on the Intonational Variation in English (IViE) speech corpus. We obtained 84.9% accuracy which is about 10% higher than previous research.

Keywords: Prominence detection, Speech clinic

1. 서 론

컴퓨터를 활용하여 영어와 같은 외국어를 학습하는 컴퓨터 보조 언어 교육(Computer-Aided Language Learning) 방식은 시간과 장소의 제약을 받지 않아 편리하고, 개인 강사가 필요하지 않아 저렴한 장점을 갖는다. 하지만 이와 같은 방법이 효과적으로 운용되기 위해서는 컴퓨터 프로그램이 학습자의 음성에서 문제점을 찾아내고 이를 학습자에게 알려줄 수 있는 능력을 갖춰야 한다.

학습자의 음성에서 찾아낼 수 있는 문제점은 여러 가지가 있을 수 있다. 우선 말의 어순이나 시제와 같은, 형식 또는 구조를 제대로 적용하지 못하여 발생하는 문법상의 오류가 있을 수 있다. 이와는 별개로 말의 리듬감이나 악센트를 제대로 활용하지 못하여 음성이 부자연스럽게 들리는 경우가 있다. 말의 리듬감이나 악센트 같은 경우는 미국식 영어에서 매우 중요한데, 참고문헌[1]에 따르면 미국인들의 경우 말의 높낮이가 계단식으로 변하지 않으면 뜻을 잘못 이해하거나 기분 나쁘게 생각할 수 있다고 한다.

주로 음성 교정기(speech clinic)는 문법적인 오류를 수정하기보다는 음성을 자연스럽게 말할 수 있도록 하는데 초점을 맞춘다. 따라서 컴퓨터 프로그램은 말의 리듬이나 악센트의 문제점을 찾아내는데 중점을 둘 필요가 있다. 특히 악센트의 위치는 참고문헌[1]에서 언급한 계단식 높낮이 변화의 기준이 되므로 정확하게 찾아낼 수 있어야 한다.

음성의 악센트는 일반적으로 고저 악센트(pitch accent)와 강세 악센트(stress accent)로 분류 할 수 있다. 이 논문에서는 음성 내에서 두 가지 악센트의 복합 작용을 통해 주변 음절에 비해서 돋보이게 들리는 특정 음절을 검출하고자 한다. 참고문헌 [2]-[4]에서는 이와 같은 음절의 특징을 다른 표현으로 '돋들림(prominence)'이라고 말하고 있다. 특히, Bagshaw는 고저 악센트나 강세 악센트를 갖고 있는 음절이 돋들림으로 지각된다고 서술하고 있다 [5].

이 사실을 바탕으로 Tamburini는 음성에서 고저 악센트와 강세 악센트를 찾아 음절의 돋들림을 판별하고 있다 [2]. 이 연구에서는 우선 각 악센트를 묘사할 수 있는 특징들을 추출한 후

1) 서울대학교 ssg@dsp.snu.ac.kr
2) 서울대학교 ksyou@dsp.snu.ac.kr
3) 서울대학교 wysung@snu.ac.kr

접수일자: 2009년 5월 1일
수정일자: 2009년 6월 8일
게재결정: 2009년 6월 16일

이렇게 찾아진 특징들을 학습하여 악센트 별로 가우시안 판별기(Gaussian discriminator)를 사용하는 감독(supervised) 방법과 추출된 특징을 점수화 하여 돋들림 여부를 판단하는 비감독(unsupervised) 방법을 제안하였다.

반면 Kochanski와 Wang은 두 악센트를 개별적으로 구분해내지 않고 관련된 특징들을 하나로 묶어 돋들림을 찾아내고 있다. 판별에 사용되는 특징들은 대부분 Tamburini가 사용한 것들과 유사하다. 다만 Kochanski는 베이즈 정리(Bayes' theorem)를 기초로 한 베이지안 분류기(Bayesian quadratic forest classifier)를 사용하였다 [3]. 그리고 Wang은 SVM(Support Vector Machine)과 간단한 평균방법을 적용하였다 [4].

돋들림 판별 정확도는 TIMIT 데이터를 쓴 Tamburini의 감독 방법이 80.73%, 비감독 방법이 80.61%이며 ICSI Switchboard 데이터와 SASO Dialog 데이터를 쓴 Wang의 감독 방법이 72.7%, 비감독 방법이 73.4%로 보고되어 있다 [2] [4]. 결과 수치들이 일반적인 음성인식기의 정확도에 비해 매우 낮게 나타나는 것을 확인할 수 있는데 이것은 돋들림 자체의 모호성 때문이다. 논문에 따르면 숙련된 경청자(listener)라 하더라도 미국 영어의 돋들림 판별 정확도가 82%에 지나지 않는다고 한다 [6]. 따라서 이 점을 고려할 때 제시된 수치는 결코 낮은 것이 아니다.

기존 연구들의 결과를 보면, Tamburini가 제안한 방식이 다른 연구들에 비해 상대적으로 높은 정확도를 가지고 있는 것을 확인 할 수 있다. 하지만 Tamburini가 제안한 방식에도 두 가지 한계점이 존재한다. 첫째, 비감독 방식에서 Tramburini는 돋들림을 찾기 위해 고저 악센트와 강세 악센트에 관련된 2개씩의 음 특징을 단순히 곱한 뒤 둘 중의 큰 값을 취해 돋들림의 상대적 점수로 사용한다. 이와 같은 방식은 각 악센트가 상호적으로 돋들림에 어떤 영향을 미치는지에 대한 분석 없이 임의적인 판별 기준을 얻어낸 것으로서 최적의 기준을 찾아냈다고 볼 수 없다. 둘째로 감독 방식의 경우 분류기를 구조화하기 위한 학습 과정을 거치면서 최적의 분류 기준을 찾는 반면에, 주변 음절의 특징을 배제하고 한 음절 내에서의 특징 벡터에 대해서만 분류를 수행함으로써 주변 음절과의 상대적인 비교가 불가능 한 단점이 있다.

본 논문에서는 위에서 언급한 두 가지 한계점을 보완하여 향상된 성능의 돋들림 검출기를 제안 하였다. 우선 학습 과정을 통해 최적의 분류 기준을 찾기 위하여 SVM (Supported Vector Machine)을 이용한 감독 방식을 적용하였다. 또한, 특징 벡터로서 각 음절의 수치화된 음 특징 뿐 아니라, 전, 후 음절 간의 음 특징 차이를 추가함으로써 주변 음절과의 상대적인 돋들림 차이를 반영 할 수 있도록 하였다.

이후의 논문 구성은 다음과 같다. 2장에서는 돋들림을 찾기 위해 필요한 각종 음 특징들과 실험에 사용된 음성 데이터에 대해 알아본다. 3장에서는 얻어진 음 특징들을 사용해 어떻게 돋들림을 판별하는지에 대해 살펴볼 것이며 추가적으로 돋들림 검출기를 음성 교정에 적용하는 방법을 기술한다. 4장에서는 돋들림 판별 알고리즘의 성능을 측정하고 분석을 할 것이고 마지막으로 5장에서는 논문의 결론을 맺는다.

2. 음 특징 추출

이 장에서는 음성의 돋들림을 판별하기위해 사용하는 음성 특징들을 추출하는 방법에 대해 설명한다. 서론에서 언급한 바와 같이 음성의 돋들림은 강세 악센트와 고저 악센트로 구성된다. 전자는 음절의 길이와 모음의 중-고 주파수 대역의 에너지와 연관이 있으며, 후자는 에너지와 음조의 움직임과 관련이 있다 [2]. 그러므로 음성의 돋들림 판별을 위해서는 입력 음성에 대한 음절의 길이, 전체에너지와 특정대역의 에너지, 음조의 움직임에 대한 데이터가 필요하다. 본 논문에서는 이 음성특징들과 함께 참고문헌 [4]에서 사용한 주파수-시간 상관관계(correlation)를 추가로 사용하여 돋들림을 판별한다.

2.1 음절 추출

Bagshaw의 돋들림 정의에 따르면 음성의 돋들림은 음절에 걸쳐있는 특징이다 [5]. 따라서 돋들림 검출에 사용되는 음 특징들은 모두 음절 단위로 얻어져야 한다. 돋들림 판별의 정확도를 높이기 위해서는 우선적으로 음절을 정확하게 추출해야 한다. 본 논문에서는 모음 에너지 또는 음성 인식기를 이용하여 음절을 추출하였다.

2.1.1 에너지를 이용한 음절 추출

본 논문에서는 참고문헌[7]에서 제시한 300-900Hz 주파수의 에너지를 이용한 음절 분리 방법을 사용한다.

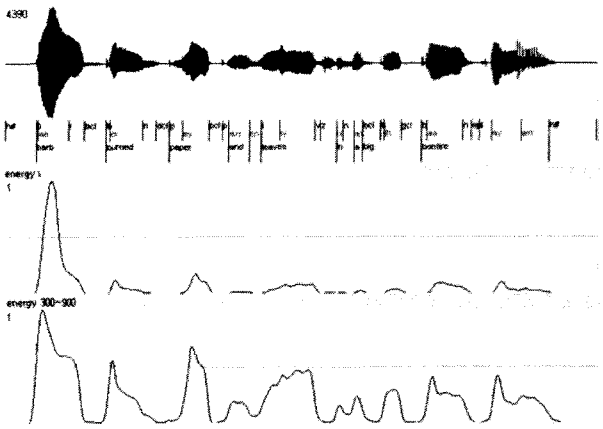

그림 1. 음절과 음성의 에너지와의 관계
Figure 1. Correlation between syllables and corresponding energy

<그림1>에 나타나듯, 300-900Hz의 음성 에너지가 전체 에너지에 비해 실제 음절의 경계와 큰 상관관계를 가짐을 볼 수 있다. 음절 추출 과정은 다음과 같다.

우선 FIR(Finite Impulse Response) 대역통과필터를 사용해 음성에서 300 - 900Hz 대역의 성분을 추출한 후 식 (1)을 사용하여 프레임 별로 에너지를 구한다. 이때 30ms의 Hamming 윈도우를 사용하여 프레임 경계의 영향을 줄인다.

$$E_{vowel}(n) = \sqrt{\frac{1}{L}\sum_{m=0}^{L-1} s(nK+m)^2 w(m)} \quad (1)$$

여기서 $s(m)$은 음성 샘플, $w(m)$은 Hamming 윈도우이다. L은 윈도우의 길이로 16kHz 음성에서는 480을 나타낸다. 음 특징의 단위가 10ms이므로 K의 값은 160이다.

음성의 300-900Hz 에너지를 추출한 후에는 참고문헌[8]에서 제시하는 수정 convex-hull 알고리즘을 이용하여 음절의 중심(syllable nuclei)을 분리하고 음절의 길이를 구한다.

300-900Hz 에너지 곡선에 convex-hull 알고리즘을 적용하기 위해서는 T_a, T_b, T_c의 총 3가지의 문턱값이 필요하다. T_a는 에너지 곡선과의 차이를 비교하여 convex-hull을 분리하기 위한 값이다. T_b는 분리된 convex-hull 내의 최대값을 음성내의 최대값과 비교해 전체적으로 낮은 에너지 구간을 없애는데 사용된다. T_c는 분리된 구간 양 끝단의 낮은 에너지 부분을 없애는데 사용된다.

T_a를 적용해 분리된 구간 중 에너지가 낮은 구간은 자음이나 잡음에 해당하므로 T_b를 적용해 이를 제거한다. 이 과정을 거치면 모음 혹은 공명자음 (sonorant consonant)만 남게 되는데, 공명 자음의 경우 모음과 구분하기가 힘들기 때문에 따로 제거하지는 않는다. 마지막으로 T_c를 적용하여 남은 구간들의 각 양쪽 끝에서 에너지가 일정 수준 이하인 지점들을 제거한다. 이는 윈도우가 자음과 모음의 경계에 걸쳐서 에너지가 높게 나와 모음구간으로 포함되는 것을 방지하기 위함이다. 구간 양 끝단의 처리가 끝나고 나면 각 구역의 길이를 쉽게 구할 수 있다. 본 논문에서는 실험적으로 얻어진 0.05, 0.05, 0.4를 각각 T_a, T_b, T_c 의 값으로 사용하였다.

<그림2>는 샘플 음성에 대해 구한 300-900Hz 에너지 곡선과 모음 구간을 보여준다. 여기서 우리는 모음 경계가 실제 그래프의 극소점에서 어느 정도 떨어져 있음을 확인할 수 있다.

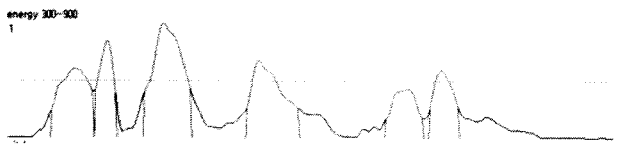

그림 2. 찾아낸 음절의 중심 영역
Figure 2. Founded syllable nucleus regions

2.1.2 음성인식기를 이용한 음절 추출

음성인식기를 이용하면 음성의 음소 단위 인식 결과를 얻어내어, 모음의 경계를 보다 정확하게 찾아낼 수 있다. 이 방식은 에너지를 이용한 방법에 비해 구현의 복잡도가 크게 증가하지만 공명자음과 모음의 구분이 가능하기 때문에 추출 성능의 향상을 기대할 수 있다. 하지만 참고문헌[8]에서 보이듯 언어 모델을 적용하지 않은 음소 음성인식기의 경우 음소 음성인식기의 낮은 인식 성능으로 인하여 음절 추출에 성능 향상에 한계가 있다. 그러므로 본 논문에서는 언어모델을 적용한 음성인식기를 사용하여 성능향상을 얻고자 하였다. 실험에는 문맥 종속 은닉 마르코프 모델(context dependent hidden Markov model) 기반의 음성인식기를 사용하였으며, 음향 모델은 Wall Street Journal 음성 데이터를 이용하여 학습된 48,000개의 가우시안으로 이루어져있다. 에너지와 음성인식기의 두 방법을 각각 적용한 돋들림 판별기의 성능 비교는 4장에서 다루었다.

2.2 에너지

음절의 전체 에너지를 얻어내기 위하여 음성의 각 프레임별 에너지 값을 다음 식 (2)와 같이 계산하였다.

$$E(i) = \sqrt{\frac{1}{N}\sum_{j=Ki}^{Ki+N-1} x_j^2}, \quad i = 0, 1, 2, \ldots \quad (2)$$

N은 윈도우의 크기로 한 프레임에 들어가는 음성 샘플의 개수이며, K는 윈도우의 샘플 이동수, i는 프레임 번호를 나타낸다. (본 논문에서 N=480, K=160 이 사용되었다.)

2.3 음조

본 논문에서는 음조를 구하기 위해 2.1절에서 구한 음절 중심을 이용한다. 무성음 구간에서는 음조가 나타나지 않지만 음절 중심은 유성음에 해당하는 모음이나 공명자음이라서 음조 성분이 존재하기 때문이다. 이렇게 음조를 처음 구하고 나면 원래 값의 두 배 혹은 절반이 되는 문제가 발생한다. 이를 보정해 주고 잡음으로 인한 급격한 변화를 없애기 위해 평활화 과정을 거친다. 다음 장에서 음조를 점수화하기 위해서는 음조가 존재하지 않는 무성음 구간에도 음조 값을 할당 하여야 한다. 이를 위해 참고문헌[9]에서 제시된 선형 보간법을 이용, 주변 음조에 의해 결정된 추정값을 구한다.

2.4 300-2200Hz 에너지

돋들림을 구분하는데 있어서 저 주파수대역이나 고 주파수 대역은 큰 영향을 미치지 못하는 반면 300-2200Hz 대역 내의 에너지는 돋들림이 존재하는 음절과 그렇지 않은 음절 사이에 두드러질만한 차이가 존재한다 [2]. 그러므로 300-2200Hz 대역에 대한 음성의 에너지를 구하여 돋들림 판별에 이용한다. 특정

대역의 에너지를 구하는 방식은 2.1.1에서 구한 방식처럼 대역통과 필터를 적용시킨 후 에너지를 계산한다.

2.5 주파수와 시간 상관

주파수와 시간의 상관관계(spectral and temporal correlation)는 돋들림 음절을 찾기 위해 참고문헌[4]에서 이용하는 특징으로 돋들림이 나타나는 위치가 모음이며 그 모음은 상대적으로 길게 발음된다는 이론을 바탕으로 두고 있다.

본 논문에서는 이러한 모음을 찾기 위해 각 모음이 고유한 대역에 에너지가 집중되어 있다는 사실을 이용한다. 모음 구간 내에서는 해당 대역들의 에너지 값이 크므로 주파수 상에서 상관 연산을 통해 모음의 위치를 찾을 수 있다. 길게 발음되는 모음의 경우 모음이 일정 시간 변하지 않고 지속 되므로 시간상에서 상관 연산을 통해 극대점을 구하고, 이를 통해 원하는 모음을 찾을 수 있다.

상관 값을 구하기 위해 참고문헌 [4]에서 제시된 방법을 사용한다. 우선 음성의 주파수 대역을 19개로 나누어 각 대역마다 에너지를 구한 뒤 가장 에너지가 높은 12개의 대역을 선택한다. 그리고 선택된 대역들 간에 주파수 상관 연산을 한 후, 시간에 따른 주파수 상관 값들에 대해 다시 시간 상관 연산을 취한다. 이 값은 다른 특징들과 동일하게 10ms 단위로 얻어진다.

3. 돋들림 판별 방법

이 장에서는 2장에서 구한 음 특징들을 바탕으로 돋들림을 찾는 방법을 서술한다.

3.1 감독 방법과 비감독 방법의 비교

이전의 돋들림 검출 방법에 대한 연구들은 검출 방식에 따라 크게 감독 방법과 비감독 방법으로 나눌 수 있다. 두 방법 간의 가장 큰 차이점은 검출을 위해 사전에 많은 데이터를 이용, 훈련을 시켜 분류 모델을 만들어야 하는지의 여부다.

감독 방법의 예로는 참고문헌[2]에서 사용한 가우시안 판별기를 들 수 있다. 이는 두 악센트와 관련된 특징들이 돋들림 여부에 따라 각각 다른 2차원 가우시안 분포를 가지고 있다는 실험 결과를 바탕으로 적용되었다. 하지만 가우시안 판별기를 적용하기 위해서는 돋들림 음절과 비돋들림 음절에 대한 각각의 훈련과정이 필요하다.

비감독 방법의 예로는 참고문헌[2] [4]에서 제안한 방식을 들 수 있다. 두 방법 모두 음 특징들과 돋들림 음절 간의 연관성을 조사하여, 돋들림의 정도를 점수화 하는 함수를 제시하였다. 각 평가함수는 식 (3)과 식 (4)에 나타나 있다. 평가 함수를 이용하면 임의의 입력 음성에 대해 돋들림의 정도를 바로 얻어낼 수 있으므로, 많은 데이터에 대해 사전에 훈련 과정을 거칠 필요가 없으며 돋들림 판단에 있어서도 많은 시간을 필요치 않는다. 따라서 비감독 방법은 비슷한 리듬 특성만 가지고 있다면 영어가 아닌 다른 언어에도 적용이 가능한 추가적 이점까지 갖고 있다.

$$prom^i = \max\left\{energy^i_{300-2200} \bullet duration^i,\ energy^i_{all} \bullet pitch_{movement}\right\} \quad (3)$$

$$PS = \frac{syl\,dur\,score + spec\,score + pitch\,max\,score}{3} \quad (4)$$

비감독 방식은 평가 함수의 값에 의해 직접적으로 돋들림 여부가 결정되므로 최적의 평가 함수를 찾는 것이 매우 중요하다. 일반적으로 다양한 화자와 환경에서 사용되는 돋들림 검출의 특성상 모든 경우에 대해 최적의 성능을 보이는 평가 함수를 찾기란 쉽지 않다. 따라서 참고문헌[4]에서는 비감독 방식이 감독 방식에 비해서 낮은 성능을 보이는 것을 알 수 있다.

하지만 참고문헌[2]에서는 비감독 방식이 감독 방식과 유사한 정확도를 보인다. 이 연구에서는 식 (3)으로 구해지는 돋들림 점수에 대하여 단순히 문턱 값과의 비교뿐 아니라 좌우 음절과의 비교로 상대적인 돋들림 여부까지 고려하여 높은 성능의 비감독 검출기를 개발하였다. 이는 돋들림이 주변 음절과의 차이에 큰 영향을 받는 다는 것을 의미한다.

따라서 본 논문에서는 감독 방식의 검출 방식을 사용하여 최적의 판단 기준을 적용하는 한편, 주변 음절과의 음 특징 차이도 함께 고려하는 돋들림 검출기를 제안한다.

3.2 음 특징 점수화

2장에서 설명된 다섯 가지 음 특징들을 돋들림 검출기에 사용하기 위해서는 각 특징을 점수(score)화할 필요가 있다. Sluijter와 van Heuven의 결과에 따르면 돋들림 음절은 긴 지속시간을 갖고 있으며 중간대역의 주파수에서 큰 에너지를 갖고 있다 [10]. 또한 2.5절에서 살펴봤듯이 돋들림 음절은 높은 주파수와 시간 상관 값을 가지고 있다.

따라서 에너지, 300-2200Hz 에너지, 주파수와 시간 상관값에 대해서는 각 음절 구간별로 최대값을 구해 그 값을 점수화 하고 음절 중심 길이의 경우 각 모음 구간의 길이를 구해 그 구간의 점수로 할당한다.

한편, 일반적으로 돋들림이 나타나는 부분에서 음조가 증가하다가 감소하는 모습을 볼 수 있는데, 그 안에서도 음조가 증가하고 있는 구간이 좀 더 돋들림과 연관성이 높다고 한다[2]. 따라서 음조를 점수화하기 위해 음조 증가의 양과 그 지속시간을 고려하는 방법을 사용하였다.

먼저 모든 프레임에 대해 음조를 구한 후에, 각 음절 중심 구간의 시작점에서부터 5개씩 값을 취해 최소 중간치 자승법(least median square)를 적용하여 근사적인 1차원 직선을 얻어낸다. 획득한 직선의 기울기가 양이라면 음의 기울기가 나올 때까지 계속해서 다음 직선을 얻어낸다. 음의 기울기를 얻었다면 이

전에 양의 기울기가 유지되었던 구간에서 음조의 변화량을 추출한다.

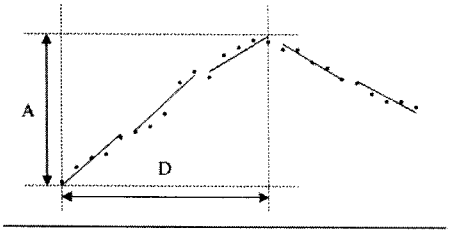

그림 3. 음조 곡선의 부분적 기울기 변화
Figure 3. Partial slope variation of a pitch contour

<그림3>과 같이 음조 증가 구간의 길이를 D, 음조 증가량을 A, 구간의 평균 음조를 P, 음성 전체에서 최대 음조를 P_m이라고 할 때 점수는 다음과 같이 계산한다.

$$score = \frac{D \cdot A \cdot P}{P_m} \tag{5}$$

3.3 SVM을 이용한 돋들림 검출

본 논문에서 제안된 돋들림 검출기는 돋들림 분류방식으로 패턴 인식에서 많이 사용되는 SVM을 이용하였다 [11] [12]. SVM 구현은 RBF(Radial Basis Function)를 커널로 사용하는 LIBSVM을 이용하였다 [13]. 학습 자료에서 각 음절의 5가지 음 특징을 점수화하여 벡터로 만들고, 이를 돋들림 여부에 따라 +1/-1로 표기하여 SVM 학습을 수행한다. 학습 자료의 모든 음절에 대해 학습이 끝난 후, 돋들림 검출기에서는 판별하고자 하는 음절의 음 특징을 추출하여 SVM모델과의 비교검증을 통해 판별 결과를 얻어낸다.

3.1에서 설명한 바와 같이, 주변 음절과의 상대적인 차이는 돋들림 여부에 큰 영향을 미친다. 참고문헌[2]의 비감독 방식에서는 돋들림 여부를 판단할 때, 식 (3)을 통해 얻어진 음절의 돋들림 점수가 주변의 다른 음절들의 점수들이 비해 돌출되어 보이는지를 확인한다. k번째 음절이 k-1번째 음절과 k+1번째 음절에 둘러싸여 있을 때, k번째 음절의 값과 좌우 음절의 값을 비교하여 두 값이 모두 k번째 음절의 85% 크기보다 작은 경우 k번째 음절을 돋들림 음절로 판단한다. 만일 인접한 음절이 k번째 음절보다 크다면 k번째 음절은 돋들림이 없는 것으로 간주된다. 또 다른 경우로 인접 음절이 k번째 음절보다는 작지만 85%의 크기는 넘는 경우 이 음절을 제외하고 다음 음절과 값을 비교하게 된다. 추가적으로 가장 큰 돋들림 값의 70% 이상에 해당하는 값을 갖고 있는 음절 또한 돋들림이 있다고 표시한다.

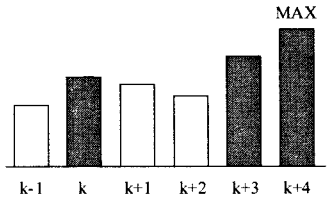

그림 4. 돋들림 점수값에서 돋들림 위치를 판별하는 예
Figure 4. Example of the prominence detection from prominence score

예를 들어 <그림4>와 같이 6개의 돋들림 값이 있을 때 k-1번째 값은 k번째 값의 85%가 되지 않으나 k+1은 초과하므로 다음 값인 k+2번째와 비교된다. 그런데 k+2번째 값이 k번째 값의 85% 미만이므로 k번째 돋들림 값은 실제 돋들림 위치로 판단되는 것이다. 그리고 k+3번째 값의 경우 k+4번째 값보다 작으므로 돋들림 위치가 될 수 없지만 최대값인 k+4의 70% 보다 큰 값을 갖고 있으므로 역시 돋들림 위치가 된다.

주변 음절과의 비교를 감독 방식에서 사용하기 위해 본 논문에서는 입력 데이터에 전처리 과정을 추가하여 좌우 음절과의 음 특징 차이를 얻어내고, 이를 SVM 모델의 학습과 검출에 적용하였다. 또한 좌우 인접한 하나의 음절과의 비교뿐 아니라, 2개 이상의 음절과 비교한 값을 특징으로 추가하여 주변 음절과의 상대적인 비교 효과를 극대화 시킨다. 비교 대상의 범위에 따른 정확도 차이 분석을 위해 4장에서는 좌우 비교하는 음절의 수를 증가시켜가며 정확도가 어떻게 변하는지에 대한 실험 결과를 제시할 것이다.

4. 실험 결과 및 분석

실험에 사용된 데이터로 IViE 음성 데이터의 'sentence' 그룹에 포함된 문장들 중에서 음성인식에 의한 음절 분리가 상대적으로 정확한 395문장을 선택하였다. 이중에서 SVM의 훈련 데이터로 사용되는 문장은 'Belfast', 'Cambridge', 'Dublin' 지역의 180문장이며 테스트에 사용되는 문장은 'Bradford', 'Leeds', 'London', 'Newcastle' 지역의 215문장이다. 훈련 데이터는 1137개의 음절을 갖고 있으며 이중에서 돋들림이 표시된 음절은 405개, 그렇지 않은 음절은 732개이다. 테스트 데이터는 1369개의 음절을 갖고 있으며 돋들림 음절은 480개, 비돋들림 음절은 889개이다.

검출기의 구성에 따른 성능 비교는 다음과 같은 세 가지 식을 바탕으로 하였다.

$$\text{정확도} = \frac{TP+TN}{TP+TN+FT+FN} \times 100 \tag{6}$$

$$\text{정밀도} = \frac{TP}{TP+FP} \times 100 \tag{7}$$

$$\text{회상율} = \frac{TP}{TP+FN} \times 100 \quad (8)$$

여기서 TP, TN, FT, FN은 각각 진실긍정, 진실부정, 허위긍정, 허위부정을 나타낸다. 진실긍정은 돋들림 음절을 그렇다고 판단한 경우, 진실부정은 비돋들림 음절을 그렇다고 판단한 경우, 허위긍정은 비돋들림 음절을 그렇지 않다고 판단한 경우, 허위부정은 돋들림 음절을 그렇지 않다고 판단한 경우이다. 따라서 정확도는 돋들림과 비돋들림을 모두 맞출 확률, 정밀도는 검출기가 돋들림으로 판단한 것 중 실제로 그러한 것의 확률, 회상율은 총 돋들림 음절 중에서 검출기가 제대로 찾은 것의 비율이다. 참고문헌[2]-[4]와 비교를 하기 위해서는 세 논문에서 공통적으로 쓴 측정 방법인 정확도를 사용할 것이다.

첫째로, 본 논문에서 제안된 주변 음절 특징 차 추가에 의한 성능 향상을 분석하기 위하여 비교 대상 주변 음절 수를 증가시켜 가면서 성능 차이를 분석하여 <표1>에 나타내었다. 표에서 비교 음절의 개수는 좌/우 각 방향의 음절 수를 의미하므로 음절의 개수가 4라는 것은 총 8개의 특징이 추가된 것을 의미한다.

표 1. 음절 수 변화에 따른 성능 비교
Table 1. Performance comparison with varying a number of syllables

비교 음절 수	측정 기준	인식기 사용	모음 에너지 사용
0	정확도	66.69	**70.49**
	정밀도	53.11	56.15
	회상율	**42.71**	24.59
1	정확도	**77.28**	76.84
	정밀도	**70.46**	68.33
	회상율	60.63	48.01
2	정확도	**83.93**	82.32
	정밀도	**80.52**	78.82
	회상율	**71.46**	59.25
3	정확도	84.37	82.83
	정밀도	81.37	77.27
	회상율	**71.88**	63.70
4	정확도	84.88	83.13
	정밀도	80.13	76.34
	회상율	**75.63**	66.51
5	정확도	83.35	83.78
	정밀도	77.27	77.19
	회상율	**74.38**	68.15
6	정확도	84.37	84.15
	정밀도	79.42	78.23
	회상율	**74.79**	68.15

<표1>에서 비교 음절수에 따라 돋들림 검출 성능이 크게 변하는 것을 알 수 있다. 비교 음절이 없는 경우와 비교했을 때 비교 음절이 하나 추가됨으로써 정확도/정밀도/회상도 모두 크게 향상되는 것을 볼 수 있다. 하지만 비교 음절 수 증가에 따라 지속적으로 성능 향상을 얻어낼 수는 없었다. 예를 들어 5개의 음절과의 비교 특징을 사용한 경우는 4개를 사용한 경우에 비해 성능이 오히려 저하되는 것을 볼 수 있었다. 본 실험에서는 4개의 비교 음절수를 갖는 경우에 가장 최적의 결과가 나타난다는 것을 알 수 있다.

음절 추출 성능의 영향을 분석하기 위해 모든 실험 결과는 음절 추출 방법으로 에너지를 사용한 경우와 음성인식기를 사용한 경우 두 가지에 대해 수행하였다.

인식기를 사용했을 때와 그렇지 않은 경우를 비교해 보면 인식기를 사용했을 경우의 성능이 더 뛰어남을 확인할 수 있다. 가장 큰 원인으로 추측되는 것은 모음 에너지를 사용한 방법이 모음과 붙어있는 공명자음을 정확하게 분리하지 못해, 모음 구간을 실제보다 길게 잡게 되어 검출 성능을 저하시키는 것이다. 실제로 모음 에너지로 구한 구간이 음성인식기로 구한 구간을 많게는 3개 이상 포함하는 경우가 흔하게 나타났다.

두 번째 실험으로 각 음 특징들이 돋들림 음절을 판단하는데 얼마나 기여하는지 알아보기 위하여 5가지 음 특징 중에 한 가지만 사용하였을 때의 돋들림 검출 성능을 분석하였다. 훈련과 테스트 과정에 모두 한 종류의 음 특징을 사용 하였으며 비교하는 음절의 개수를 0에서 최적 값인 4까지 바꿔가며 실험을 하였다.

표 2. 각 음 특징이 검출기 성능에 미치는 영향
Table 2. Performance related to a each acoustic feature

비교 음절 수	측정 기준	모음 길이	대역 에너지	음조	에너지	주파수와 시간 상관 값
0	정확도	64.57	66.40	64.94	67.35	**70.12**
	정밀도	47.71	55.05	n/a	58.92	**65.11**
	회상율	10.83	22.71	0	22.71	**31.88**
1	정확도	72.68	72.97	65.89	**75.31**	74.07
	정밀도	65.96	67.97	60.00	**72.19**	63.68
	회상율	45.63	43.33	8.125	48.13	**60.63**
2	정확도	78.74	79.11	66.25	79.99	**82.47**
	정밀도	75.89	76.80	61.84	78.93	**81.41**
	회상율	57.71	57.92	9.79	58.54	**64.79**
3	정확도	80.42	79.47	65.52	80.20	**82.91**
	정밀도	75.73	73.09	52.6	72.97	**76.86**
	회상율	65.00	65.63	16.88	69.17	**73.33**
4	정확도	81.15	80.79	67.20	81.30	**84.73**
	정밀도	**81.36**	72.75	57.28	74.67	79.91
	회상율	60.00	72.29	25.42	70.63	**75.42**

<표2>의 결과를 확인하면 비록 한 종류의 음 특징을 사용하지만 여전히 비교 음절의 개수가 증가함에 따라 성능이 급격하게 향상됨을 확인할 수 있다. 하지만 유일하게 음조의 변화는 성능향상이 미미한데 이는 IViE의 음성 데이터가 급격한 음조 변화를 가지지 않기 때문이다. 따라서 음조 점수가 전혀 나타나지 않는 음절이 상당히 많았으며 제대로 된 판별 기준으로 작용할 수 없었다.

나머지 네 가지 특징들은 대동소이한 모습을 보이나 '주파수와 시간 상관 값'이 가장 뛰어난 성능을 보여주고 있다. 심지어는 5개의 음 특징을 모두 사용한 경우와도 유사한 성능을 보여주고 있는데 이를 통해 주파수와 시간 상관 값이 검출기의 성능을 가장 크게 좌우한다고 판단 할 수 있겠다.

마지막으로 다섯 가지 특징 가운데 하나를 제거했을 때의 성능을 비교해 보았다. 두 번째 실험에서 음조가 돋들림 판별에 낮은 성능을 보였기 때문에, 오히려 성능을 저하시키지는 않는지 확인해 보기 위함이다. 실험 결과는 <표3>에 나타나 있다.

표 3. 각 음 특징을 제거했을 때의 성능 비교
Table 3. Performance comparison with excluding a feature

비교 음절 수	측정 기준	모음 길이	대역 에너지	음조	에너지	주파수와 시간 상관 값
0	정확도	67.49	**70.34**	65.38	68.08	65.81
	정밀도	55.63	**62.33**	51.06	56.13	51.55
	회상율	36.04	38.96	30.00	41.04	**41.46**
1	정확도	75.75	76.70	**78.82**	76.63	75.60
	정밀도	66.67	68.76	**71.11**	69.32	69.01
	회상율	61.67	61.46	**66.66**	59.79	55.21
2	정확도	82.69	82.98	82.98	**83.35**	79.25
	정밀도	77.30	78.52	80.20	**80.88**	76.34
	회상율	**71.67**	70.83	68.33	68.75	59.17
3	정확도	83.20	84.51	**84.95**	83.86	81.59
	정밀도	78.28	78.88	**81.28**	78.09	76.89
	회상율	72.08	**76.25**	74.17	75.00	67.92
4	정확도	84.37	83.93	84.44	**84.88**	82.76
	정밀도	80.23	79.68	79.47	**81.09**	78.11
	회상율	73.54	72.71	**75.00**	74.17	70.63

실험 결과를 보면, 음조가 포함되거나 제거 되었을 경우의 성능이 <표1>의 결과와 거의 유사하다. 이는 음조가 실제로 성능의 향상이나 저하에 아무런 역할을 하고 있지 않다는 것을 의미한다.

나머지 네 가지 특징 중에서 '주파수와 시간 상관 값'은 <표2>에서 살펴봤듯 돋들림에 가장 큰 영향을 미치므로, 이 실험에서도 이 값이 빠졌을 때 많게는 3% 이상의 가장 큰 성능저하가 나타남을 확인할 수 있다.

표 4. 타 연구 결과와의 비교
Table 4. Performance comparison

	정확도	정밀도	회상율
[2] 감독방법	80.7	70.0	**77.0**
[2] 비감독방법	80.6	75.3	64.3
[3]	78.6	n/a	n/a
[4] 감독방법	76.2	**82.1**	73.4
[4] 비감독방법	73.4	80.0	70.0
제안된 방법 0음절	66.7	53.1	42.7
제안된 방법 4음절	**84.9**	80.1	75.6

본 논문에서 제안된 돋들림 검출기의 성능을 기존 연구들의 성능 결과와 비교하여 <표4>에 나타내었다. 참고문헌[3]을 제외하고는 사용한 음성 데이터가 다르기 때문에 직접적인 비교는 불가능하다. 좌우 4음절의 특징 차를 적용한 검출기의 경우에 기존 연구들에 비해 상당히 높은 정확도 수치를 가진다. 정밀도와 회상율도 상위에 분포하고 있어 제안된 방법이 효율적으로 돋들림 음절을 찾아낼 수 있음을 알 수 있다.

5. 결 론

본 논문에서는 영어 발음 교정에 사용될 수 있는 돋들림 검출 알고리즘을 제안하였다. 보다 정확한 검출 성능을 얻어내기 위하여 각 음절의 음 특징 뿐 아니라 좌우 음절과의 음 특징 차를 고려하여 감독 방식의 검출기의 특징 벡터로 사용하였다. 또한 보다 정확한 음절 추출을 위하여 음성인식기를 사용하였다. 본 논문에서 제안된 돋들림 검출기의 검출 정확도는 최고 84.9%로서 기존 연구들에 비해 최대 10%이상 나은 성능을 얻어내었다. 추후 더 향상된 검출 성능을 얻어내기 위해서는 추출된 음절 구간에 대한 후처리를 통하여 공명자음과 모음을 더 정확하게 구분할 수 있도록 하는 연구가 필요하다.

감사의 글

이 연구에 참여한 연구자는 「2단계 BK21사업」의 지원비를 받았음.

참 고 문 헌

[1] Cook, A., Jeon, C. H. (2007). *AAT-American Accent Training.* Willbook, pp. 18-37.
(앤 쿡 · 전창훈, (2007). AAT-미국식 영어발음 집중훈련 워크북, 윌북, 2007, pp.18-37)

[2] Tamburini, F., Caini, C. (2005). "An automatic system for detecting prosodic prominence in American English continuous speech," *International Journal of Speech Technology*, Vol. 8, pp. 33-44.

[3] Kochanski, G., Grabe, E., Coleman, J. et al. (2005). "Loudness predicts prominence: fundamental frequency lends little," *J. Acoust. Soc. Amer.*, Vol. 118, No. 2, pp. 1038-1054, Aug.

[4] Wang, D., Narayanan, S. (2007). "An acoustic measure for word prominence in spontaneous speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 15, No. 2, pp. 690-701, Feb.

[5] Bagshaw, P. C. (1994). *Automatic Prosodic Analysis for Computer-Aided Pronunciation Teaching*, Ph. D. Dissertation, University of Edinburgh.

[6] Jenkin, K. L., Scordilis, M. S. (1996). "Development and comparison of three syllable stress classifiers," *ICSLP '96*

Proceedings, Philadelphia, pp. 1457-1460.
[7] Howitt, A. W. (2000). *Automatic Syllable Detection for Vowel Landmarks*, Ph. D. Dissertation, MIT.
[8] Xie, Z., Niyogi, P. (2006). "Robust acoustic-based syllable detection," *In Proc. of ICSLP*.
[9] Taylor, P. (2000). "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, Vol. 107, pp. 1697-1714.
[10] Sluijter, A., van Heuven, V. (1996). "Acoustic correlates of linguistic stress and accent in Dutch and American English," *ICSLP'96 Proceedings*, Philadelphia, pp. 630-633.
[11] Kim, Y. D., Kim, K. H., Song, S. H. (2005). "Comparison of boosting and SVM," *Journal of the Korean Data & Information Science Society*, Vol. 16, No. 4, pp. 999-1012.
[12] Hsu, C. W., Chang, C. C., Lin, C. J. (2008). "A practical guide to support vector classification," http://www.csie.ntu.edu.tw/~cjlin.
[13] Chang, C. C., Lin, C. J. (2008). "LIBSVM," http://www.csie.ntu.edu. tw/~cjlin.

• 심성건 **(Shim, Sunggeon)**
서울대학교 전기,컴퓨터공학부
서울시 관악구 신림9동 서울대학교 공과대학
Tel: 02-880-9372 Fax: 02-882-4656
Email: ssg@dsp.snu.ac.kr

• 유기선 **(You, Kisun)**
서울대학교 전기,컴퓨터공학부
서울시 관악구 신림9동 서울대학교 공과대학
Tel: 02-880-9372 Fax: 02-882-4656
Email: ksyou@dsp.snu.ac.kr

• 성원용 **(Sung, Wonyong)** 교신저자
서울대학교 전기,컴퓨터공학부
서울시 관악구 신림9동 서울대학교 공과대학
Tel: 02-880-9372 Fax: 02-882-4656
Email: wysung@snu.ac.kr
1989~현재 전기,컴퓨터공학과 교수