강세/비강세 모델을 이용한 강세 판정 방법

김우일*, 고훈**, 고한석*
*고려대학교 전자공학과, **(주)보이스미디어텍

Model based Stress Decision Method

Wooil Kim*, Hoon Koh**, Hanseok Ko*
*Dept. of Electronics Engineering, Korea Univ., **VoiceMediaTech Inc.
wikim@ispl.korea.ac.kr

요 약 문

본 논문에서는 일반적인 강세 판정법이 갖는 단점을 보완하기 위하여 모델을 기반으로 하는 강세 판정 방법을 제안한다. 기존의 강세 판정법은 기준값과의 절대적인 비교에 의해 강세를 판정하게 되므로 발음 환경에 따라 불안정한 성능을 나타낸다. 제안하는 방법은 강세/비강세 모델을 적용한 후보들에 대해 상대적인 비교값으로 강세를 판정한다. 소량의 강세 음성 데이터베이스로부터 강세/비강세 모델을 훈련하기 위해 적응 훈련 기법을 사용하였다. 실험 결과 76.53%의 판정 성공률을 나타내었으며, 이는 제안한 방법이 강세 자동 판정 방법으로 이용 가능성을 보이는 결과이다.

I. 서론

음성 인식 기술에 대한 꾸준한 관심과 연구의 결과로 최근에는 영어 발음을 평가해주는 학습 장치로의 적용 사례와 실제 서비스가 등장하고 있다. 일반적인 영어 학습 장치는 음성 인식 기술을 이용하여 사용자가 발음한 음성 신호를 단어 혹은 문장의 훈련 모델에 적용한 확률 값의 높낮이에 따라 발음의 숙련 여부를 측정해주는 방법을 사용한다. 이러한 형태의 발음 학습 장치는 인터넷상에서 구현되어 서비스되거나 컴퓨터 게임 형식 등으로 출시되고 있다. 음성 인식 기술을 이용한 학습 장치 형태 외에 사용자(피험자)의 발음에서 추출된 다양한 음향학적 자질에 대한 분석과 표준 발음과의 비교 등을 통해 발음의 정확성 여부를 평가하는 연구가 음성학 분야에서 이루어져 왔다. 이러한 연구는 언어 교육, 발음 장애 치료 등을 목적으로 실험적 차원에서 이루어져 왔으나, 최근에 음성 신호 처리 기술과의 접목을 통해서 음향학적 자질 분석을 통한 자동 발음 평가 장치 구현에 관한 노력이 계속되고 있다[1].

본 논문에서는 영어 발음의 정확성 평가에 있어 주요 요소가 되는 강세(stress)의 위치를 자동으로 판정하기 위한 방법을 제안한다. II장에서는 기존의 방법에서 발견되는 문제점들을 지적하고 이를 보완할 수 있는 강세 판정 방법을 제안한다. III장에서는 강세/비강세 모델을 훈련하기 위해 사용된 Bayesian 적응 기법을 소개하고 IV장에서는 제안하는 방법을 자세히 설명한다. V장에서는 적응 기법의 성능 검증을 위한 화자 적응 실험을 보이고, 제안한 방법의 강세 판정 성능을 실험과 결과로 확인한다. 마지막으로 VI장에서 결론을 맺고 향후 연구 방향에 대해 논의한다.

II. 문제 제기

발음에서 강세는 일반적으로 피치(pitch), 발음의 세기, 지속 시간 등의 요인이 전체적인 효과로 나타내게 된다. 강세가 있는 구간에서는 피치가 높고, 세기가 크며, 지속기간이 길어진다[2]. 강세 구간을 검출하기 위한 방법으로 이와 같은 음향학적 자질을 각각 검출하여 기준 패턴과의 비교를 통해 강세 구간의 여부를 판단하는 방법이 일반적이다[1]. 하지만, 다음과 같은 이유로 정확한 강세 판정에 어려움을 겪고 있다.

첫째로 각 음향적 자질의 정규화의 어려움이다. 각 음향적 자질들은 절대적인 값이 아니고, 화자나 발음 환경 등에 따라 달라지는 값들이므로 이를 측정 단위로 사용하기 위해서는 적절하게 정규화하고 변형하여 사용하여야 한다. 하지만, 이러한 음향 자질들의 변이를 예측하기 힘들고 범위 또한 일정하지 않기 때문에 정규화에 어려움이 있고, 임의적인 방법으로 사용하는 것이 일반적이다. 피치나 포만트(Formant)의 경우, 화자 적응(Speaker Adaptation)의 한 방법으로 성도 정규화(Vocal Tract Normalization)와 같은 방법이 사용되지만, 이는 화자에 대한 어느 정도의 음성 데이터 내지 사전 정보를 가지고 있어야 한다[4].

둘째는 강세 판정에 필요한 객관적인 기준 값을 설

정하기가 힘들다는 점이다. 일반적인 강세 판정 방법에서는 추출된 음향 자질들을 동적 시간 정합(DTW, Dynamic Time Warping)과 같은 매칭 방법에 의해 비교 값을 계산하게 되고, 이 비교 값을 설정한 기준 값에 따라 강세의 유무를 판단하게 된다. 판단에 사용되는 기준 값은 다량의 데이터에 대한 통계 값으로 설정하게 되는데, 환경에 따라서 음향 자질들의 값이 변이가 심하고, 그 만큼 정규화 결과도 정확하지 않으므로 기준 값과의 절대적인 비교가 어렵게 된다. 이러한 부분은 객관적인 강세 판정에 큰 장애요소가 되며, 환경에 따라 불안정한 성능을 갖게 한다[4].

본 논문에서는 모델 기반의 평가 방법을 도입하여 절대적인 기준 값에 의한 판단이 아닌, 후보에 대한 상대적인 비교를 통해 강세의 위치를 결정하고자 한다. 각 모음에 대해서 강세를 갖는 경우와 그렇지 않은 경우에 대해 각각 음향 모델을 생성하고, 이들의 조합에 의한 발음 사전에 따른 상대적인 비교 값에 의해 강세 위치를 판정한다. 강세/비강세 모음에 대한 모델을 훈련하기 위해서는 강세를 고려한 음성에 대해 충분한 데이터베이스를 구축하는 것이 요구되는데, 이러한 조건의 데이터베이스를 충분히 구축하는 것은 힘든 일이므로 본 논문에서는 적응 훈련(adaptive training) 기법을 통해 적은 양의 데이터베이스로부터 강세/비강세 모음에 대한 음향 모델을 유도해 내고자 한다.

III. Bayesian 적응 기법

음성 인식에서의 적응 훈련이란 음향 모델에 대한 사전 확률(a prior probability) 분포를 가정함으로써 특정 화자 혹은 인식 환경 등에 적응적이고, 적합한 음향 모델을 생성하여 인식률을 향상시키는 훈련 기법을 말한다. 본 논문에서는 Bayesian 적응 기법의 하나인 사전 확률 밀도를 포함한 MAP(maximum a posteriori)을 통한 이산 확률 분포 HMM 파라미터의 적응 훈련 방법을 사용하고자 한다[5].

N 개의 상태(state) 수를 갖는 이산(discrete) 확률 분포 HMM(Hidden Markov Model)의 파라미터 벡터를 $\lambda=(\pi, A, B)$ 라 할 때 각 파라미터가 독립이라고 가정하고 각 열에 대해 독립임을 가정하면, 각 파라미터에 대한 확률 밀도는 Dirichlet 분포를 갖게 되고, 사전 확률 밀도 $g(\lambda)$는 다음과 같이 행렬 beta 확률 밀도 함수의 특수한 형태가 된다.

$$g(\lambda)=K_c\prod_{i=1}^{N}\left\{\pi_i^{\eta_i-1}\cdot\left(\prod_{j=1}^{N}a_{ij}^{\eta_{ij}-1}\right)\cdot\left(\prod_{k=1}^{K}b_{ik}^{\nu_{ik}-1}\right)\right\} \tag{3.1}$$

여기에서 K_c는 정규화 상수이고, $\eta_i, \eta_{ij}, \nu_{ik}$ 는 확률 밀도 함수 표현을 위한 파라미터이다. 관찰 벡터 열 x과 사전 확률 밀도 $g(\lambda)$을 알고 있을 때, 사전 확률 밀도를 포함한 HMM 파라미터 λ 에 대한 MAP 예측은 다음 식과 같다.

$$\lambda_{MAP}=\arg\max_{\lambda} P(x|\lambda)g(\lambda) \tag{3.2}$$

식 (3.2)을 EM(Expectation Maximization) 알고리즘을 적용하여 풀어보면 각 파라미터는 다음과 같이 예측 할 수 있다.

$$\hat{\pi}_i=\frac{e_i+\eta_i-1}{\sum_{i=1}^{N}(e_i+\eta_i-1)} \quad i=1,2,\ldots,N \tag{3.3}$$

$$\hat{a}_{ij}=\frac{c_{ij}+\eta_{ij}-1}{\sum_{j=1}^{N}(c_{ij}+\eta_{ij}-1)} \quad i,j=1,2,\ldots,N \tag{3.4}$$

$$\hat{b}_{jk}=\frac{d_{jk}+\nu_{jk}-1}{\sum_{k=1}^{K}(d_{jk}+\nu_{jk}-1)} \quad j=1,2,\ldots,N,\ k=1,2,\ldots,K \tag{3.5}$$

여기에서 e_i, c_{ij}, d_{jk} 는 각각 다음과 같다.

$$e_i=\Pr(s_1=i|x,\lambda) \tag{3.6}$$

$$c_{ij}=\sum_{t=1}^{T-1}\Pr(s_t=i, s_{t+1}=j|x,\lambda) \tag{3.7}$$

$$d_{jk}=\sum_{t:x_t\sim\nu_k}\Pr(s_t=j, x_t\sim\nu_k|x,\lambda) \tag{3.8}$$

s_t는 시간 t에서의 상태를 말하며, $x_t\sim\nu_k$ 는 관찰 벡터 x_t가 출력 심볼 ν_k 에 대응됨을 의미한다.

식 (3.2)에 의해 HMM 파라미터를 예측하기 위해서는 사전 확률 밀도 함수 $g(\lambda)$와 관련된 파라미터 $\eta_i, \eta_{ij}, \nu_{jk}$ 를 찾아내는 것이 필요하다. 이 하이퍼 파라미터는 두 가지 방법으로 찾을 수 있는데 첫 번째는 음향 모델에 대한 사전 확률을 얻을 수 있는 각각의 종류에 대한 HMM 파라미터들의 평균, 분산과 같은 통계 값으로부터 얻는 방법이다. 이러한 방법은 각 화자나 여러 환경에 대한 HMM 파라미터를 각각 가지고 있을 때 유효하다. 두 번째는 화자 독립 모델과 같은 각 화자 혹은 환경이 통합된 데이터베이스로 훈련된 모델로부터 구하는 방법이다. 즉, 통합된 훈련 데이터베이스에 대한 HMM 파라미터를 예측하는 과정에서 산출되는 $\hat{e}_i$, $\hat{c}_{ij}$, $\hat{d}_{jk}$ 를 각 훈련 단위의 훈련 token 수로

나눈 값에 1을 더하여 하이퍼 파라미터 $\eta_i, \eta_{ij}, \nu_{jk}$ 로 각각 사용한다[5]. 본 논문에서는 후자의 방법을 택하여, 강세를 고려하지 않은 음소 모델로부터 하이퍼 파라미터를 구하고, 이를 기반으로 강세를 고려한 소량의 음성 데이터베이스를 이용하여 강세/비강세 모음에 대한 HMM 모델을 유도하였다.

IV. 강세/비강세 모델을 이용한 강세 판정법

우선, 강세가 고려되지 않은 일반 음성 데이터베이스로부터 화자 독립의 음성 인식이 가능한 음소 모델을 생성한다. 강세가 고려된 음성 데이터베이스는 개개의 발음에 대해 강세가 표시된 발음 사전이 제공되게 된다. 즉, 강세가 표현된 모음과 그렇지 않은 모음이 구분되어 있다. 강세가 고려된 음성 데이터베이스와 3장에서 소개한 Bayesian 적응 기법을 이용하여 강세가 반영된 음소 모델을 생성한다. 적응 훈련에 사용하는 음향 모델의 사전 확률 밀도 함수에 필요한 하이퍼 파라미터는 이전에 훈련되어진 강세를 고려하지 않은 음소 모델로부터 얻어진다. 최종적으로 각 모음에 대해서 강세/비강세 쌍의 2중 모델을 갖게 된다.

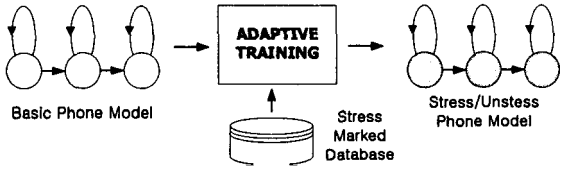

그림1. 적응 훈련에 의한 강세/비강세 모델 생성

발음 평가 장치는 사용자(피험자)에게 특정 단어나 문장에 해당하는 발음을 요구하는 형태이므로, 사용자가 어떤 발음을 할지 미리 알 수 있다. 사용자가 주어진 단어를 발음하면, 이를 평가하기 위해 단어에 해당하는 기본 음소열을 구성한다. 기본 음소열의 각각의 음절에 강세 모음을 하나씩 포함하게 하여, 총 음절 수와 일치하는 후보 음소열을 생성한다. 입력 음성 신호에 각 후보 음소열을 발음 사전으로 하는 HMM 모델을 대응시키고, Viterbi 탐색을 이용하여 각 후보에 대한 확률 값을 계산한다. 얻어진 확률 값 중에서 최대 값에 해당하는 후보 음소열이 갖는 음절의 강세 위치에 따라 입력 음성 신호의 강세 위치를 결정한다.

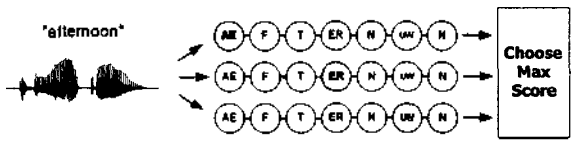

그림2. 강세/비강세 모델을 이용한 강세 판정 방법

V. 실험 및 결과

영어 발음의 강세 평가를 위해서는 영어 발음의 데이터베이스가 필요하나, 사용이 여의치 않아 본 실험에는 KAIST에서 제작한 한국어 음성 데이터베이스인 PBW452 데이터를 사용하였다. 우선 화자 독립 음소 모델 기반의 고립 단어 인식기를 구성하고, 적응 훈련 방법의 성능을 검증하기 위해 화자 적응 테스트를 시행하였다. 본 논문의 실험의 기본이 되는 화자 독립 고립 단어 인식기의 구성과 성능은 표 1과 같다.

표1. 화자독립 고립단어 인식기 [4]

어휘수	100 단어
HMM 모델	45개 음소 모델, Left-to-Right 모델, 3상태 이산 확률 분포
특징 추출	16kHz, 16bit 샘플링, MFCC(Mel Frequency Cepstral Coefficients) 38차원 특징 벡터
벡터 양자화	LBG 알고리즘, 256 크기 코드북
훈련 과정	Baum-Welch 알고리즘
인식 과정	Viterbi 디코딩
데이터베이스	훈련 - 10명의 성인 남자, 총 4520 단어 인식 - 훈련에 참여하지 않은 5명의 성인 남자, 총 500 단어
인식률	92.60%

적응 훈련 기법의 성능을 검증하기 위하여 다음과 같이 다양한 방법으로 모델을 훈련하였다. 적응시키는 화자는 기본 화자 독립 모델 훈련에 참여하지 않은 화자로서 PBW1 데이터는 훈련과 적응에, PBW2 데이터 중 100단어를 인식 테스트에 사용하였다.

1) SI1 : 화자 독립 모델 (10명의 4520 단어)
2) SI2 : 테스트 화자의 데이터를 포함한 화자 독립 모델 (10명 4520단어+1명 452단어)
3) SD : SI1을 초기 모델로, 테스트 화자 데이터로 훈련한 화자 종속 모델(1명 452 단어)
4) SA : SI1로부터 테스트 화자로 적응 훈련 (1명 452단어)

표2. 화자 적응 실험

훈련 방법	SI1	SI2	SD	SA
인식률 (%)	90	95	80	96

표 2의 실험 결과와 같이 화자 적응을 시켰을 때 96%로 인식 성능이 가장 좋은 것으로 나타났다. SD의 경우, 비록 독립 모델을 초기 모델로 사용하였으나 SD 모델 훈련을 위한 데이터베이스가 충분하지 못하므로 화자 독립 모델(SI1, SI2)보다 인식 성능이 월등히 저하되는 것으로 판단된다. 실험 결과로부터 본 실험에서 사용하는 적응 기법이 소량의 데이터로부터 특정 환경에 적응 훈련하여 인식 성능을 향상시키는 성능을 가짐을 입증 할 수 있었다. 이는 강세를 고려하지 않은 음소 모델로부터 소량의 강세를 고려한 음성 데이터를 가지고 적응 기법을 사용하여 강세/비강세 음소 모

델을 생성해내는 과정이 타당함을 반증하는 것이다.

표3. 강세 모델 적응 훈련을 위한 단어 세트

어두운, 뭐라고, 애초에, 수필집, 바쁘게, 이유에, 통증이, 우아한, 더이상, 제조업, 일제히, 이윽고, 배우자, 고마워, 재검토, 낮잠을, 그처럼, 아파트, 외무부, 드디어, 구호품

강세/비강세 음소 모델을 훈련하기 위해서는 강세가 들어간 음성 데이터가 필요하다. 음소 모델 훈련에 사용 가능한 데이터가 한국어 음성 데이터베이스이므로 강세가 들어간 한국어 음성 데이터를 임의로 제작하였다. 한국어의 주요 모음 7개(아, 애/에, 어, 오, 우, 으, 이)가 고르게 들어간 총 21개의 3음절 단어 세트를 구성하고, 성인 8명에게 매 단어마다 각각의 음절에 강세를 넣어 발음하도록 하여 총 504(8×3×21)개의 데이터를 수집하였다. 이 중에서 6명이 발음한 378개의 단어를 강세/비강세 모델 적응 훈련에 사용하였고, 나머지 2명이 발음한 126개의 단어 중에서 주요 모음만이 들어간 98개의 단어로 강세 판정을 실험하였다.

적응 훈련 결과 주요 모음에 대해서 강세/비강세 모델을 각각 얻을 수 있었다. 강세 판정 실험은 인식하고자 하는 단어에 해당하는 발음 사전을 각 음절에 강세 모델이 들어가도록 3개의 후보를 구성한다. 예를 들어, "어두운"이라는 발음의 강세를 평가하고자 할 때에는 "어"에 강세 모델, 나머지 "두", "운"에는 비강세 모델을 적용한 발음 사전을 첫 번째 후보, "두"에만 강세 모델을 적용한 두 번째 후보, "운"에만 강세 모델을 적용한 세 번째 발음 사전을 구성하고 입력 음성에 대해서 Viterbi 탐색을 통해 최대 확률값을 각각 구하였다. 계산한 확률값 3개 중에서 최대값을 갖는 후보에 해당하는 강세 모델의 위치를 측정한 강세의 위치로 판정하였으며 그 결과는 표 4와 같다.

표4. 강세/비강세 모델을 이용한 강세 판정 실험

테스트 데이터	훈련 데이터	비훈련 데이터
판정 성공률 (%)	95.37	76.53

제안한 방법으로 강세 위치를 판정한 결과 76.53%의 성공률을 얻었다. 비록 강세 위치 자동 판정의 가능성을 보이는 결과이지만, 만족할만한 성능을 얻지는 못했다. 기대에 미치지 못하는 이유를 다음과 같이 생각해 볼 수 있다. 우선 강세는 초 분절 자질의 하나로서 본 실험에서 설정한 강세 모음과 같이 하나의 분절음 구간으로는 충분히 반영하기가 힘든 것으로 판단된다. 즉, 모음뿐 아니라 인접한 다른 음소에 걸쳐 형성되기 때문에 이러한 영향들을 효과적으로 반영해주어야 한다는 점에서 한계가 있었던 것으로 생각된다. 강세 판정에 실패한 데이터들을 살펴 본 결과 경음이나 격음이 들어간 부분이 강하게 발음되어 강세 음절과 그 정도가 불분명한 경우가 있었는데, 성능 개선에 있어 참고해야 할 부분이다. 또한, 실험에 사용한 강세 단어가 인위적으로 발음되어진 단어이므로 자연스러운 강세 형성이 되지 않았던 것으로 관찰되었다. 분석 결과, 훈련에 사용되어진 어떤 발음들은 피치는 일정하고 세기만 커지거나, 지속 기간만 지연되는 등 정확한 강세가 표현되지 않은 데이터들이 발견되었다. 이러한 데이터들은 적응 훈련에 불필요한 요소로 작용되어 충실한 훈련 모델을 생성하는데 방해가 되었던 것으로 판단된다. 이러한 문제점은 향후 영어 데이터베이스를 사용할 경우 자연스러운 강세 현상을 반영하게 되므로 해결이 가능한 것으로 보인다.

본 실험을 통해서, 강세/비강세 모델로 강세 판정이 가능함을 확인할 수 있었고, 적응 훈련 기법을 이용하여 충분하지 못한 데이터베이스로부터 강세/비강세 모델을 만들어내는 것이 효과적임을 알 수 있었다. 제안한 방법은 후보 강세 발음 사전에 대한 상대적인 확률값으로 평가하는 과정이므로, 음향 자질을 이용한 표준 패턴과의 매칭을 통해서 강세를 판단하는 기존의 방법이 화자, 사용 환경에 따라 불안정한 성능을 보이는 단점을 보완 할 수 있을 것으로 판단된다.

VI. 결론

본 논문에서는 기존의 강세 판정 방법의 단점을 극복하기 위해서 강세/비강세 모델을 이용하여 강세의 위치를 결정하는 방법을 제안하였다. 실험 결과 76.53%의 강세 판정 성공률을 얻었고 이는 제안한 방법이 자동 강세 판정 방법으로 이용 가능성을 보여주는 결과이다. 향후에는 모음 뿐 아니라 자음에 대해서도 강세/비강세 모델을 확대하여 적용함으로써 초 분절적인 특성을 반영하는 방안을 검토 중이며, 일반적 강세 판정법인 자질을 이용한 방법과의 조합을 통해 강세 판정의 신뢰도를 높이는 연구를 계획하고 있다.

참고 문헌

[1] 구희산 외, *음성학과 음운론*, 한신문화사, 1995.

[2] 백승권 외, "영어 단어 발성시의 오류 교정 기술에 관한 연구," *한국 음성 과학회 제 8회 학술대회 논문집*, pp. 83-90, Apr. 2000.

[3] J. R. Deller, Jr, et al., Discrete-Time Processing of Speech Signals, Prentice-Hall, 1987.

[4] L. Rabiner, et al, Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[5] Qiang Huo, "Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition, " *IEEE Trans. on Speech and Audio Processing*, Vol.3, No.5, pp.334-345, Sep. 1995.