한국음향학회지 제25권 제2호 pp. 88~95 (2006)

고속 발화음에 대한 음성 인식 향상

Improvements on Speech Recognition for Fast Speech

이 기 승*
(Ki-Seung Lee*)

*건국대학교 정보통신대학 전자공학부
(접수일자: 2005년 12월 29일; 채택일자: 2005년 1월 20일)

본 논문에서는 대화체 음성에 대한 음성 인식의 성능을 향상시키기 위한 방법으로, 고속 발화음에 대해 강인한 음성 인식 방법을 제안하고 성능을 평가하였다. 제안된 기법은 입력된 음성의 속도를 정량화하여 나타내기 위한 부가적인 음성 인식 과정이 필요치 않으며, 특정 대역내의 에너지 분포를 이용하여 모음 구간을 판정하고, 단위 시간당 모음의 개수를 구하여 음성의 속도를 측정하였다. 빠른 발성음에 대한 음성 인식의 성능을 향상시키기 위해, 기존의 방법은 표준 음소 길이와 측정된 음소 길이간의 비율을 이용하여 특징 벡터를 시간축으로 확장하였다. 제안된 방법에서는 발성 속도에 따라 음성을 분류하고, 분류된 음성에 대해 서로 다른 시간축 확장 비율을 정하도록 하였다. 여기서 분류에 필요한 문턱치들과 시간축 확장 비율들은 최대 우도 방법을 이용하여 구하였다.
10자리 이동 전화 번호에 대한 음성 인식의 실험 결과, 제안된 기법에 의해 전체적으로 17.8% 오류율이 감소되는 것을 확인할 수 있었다.

핵심용어: 자동 음성 인식, 최대 우도, 발성 속도

투고분야: 음성 처리 분야 (2.5)

In this paper, a method for improving the performance of automatic speech recognition (ASR) system for conversational speech is proposed, which mainly focuses on increasing the robustness against the rapidly speaking utterances. The proposed method doesn't require an additional speech recognition task to represent speaking rate quantitatively. Energy distribution for special bands is employed to detect the vowel regions, the number of vowels per unit second is then computed as speaking rate. To improve the performance for fast speech, in the pervious methods, a sequence of the feature vectors is expanded by a given scaling factor, which is computed by a ratio between the standard phoneme duration and the measured one. However, in the method proposed herein, utterances are classified by their speaking rates, and the scaling factor is determined individually for each class. In this procedure, a maximum likelihood criterion is employed.

By the results from the ASR experiments devised for the 10-digits mobile phone number, it is confirmed that the overall error rate was reduced by 17.8% when the proposed method is employed

Keywords: Automatic Speech Recognition, Maximum likelihood, Speaking Rate

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

음성 인식 (Automatic Speech Recognition; ASR)은 일반적으로 학습 데이터 (training corpus)를 이용하여 인식하고자 하는 어휘, 문법 등에 대해 개별적인 통계 모델링 (statistical modeling)을 수행하고, 임의의 음성이 입력되었을 때 통계적으로 가장 유사한 어휘나 문법을 찾아 이를 인식된 결과로 출력하는 것이다[1]. 이 때, 입력된 음성이 학습 시 사용된 음성과 유사한 특징을 갖는 경우, 높은 인식율을 기대할 수 있으나, 학습 시 사용된 음성과 상이한 음성이 입력된다면 인식시 오류율이

책임저자: 이 기 승 (kseung@konkuk.ac.kr)
143-701 서울특별시 광진구 화양동 1번지
건국대학교 정보통신대학 전자공학부
(전화: 02-450-3489; 팩스: 02-3437-5235)

증가될 수 있다[2-3]. 학습 데이터와 차이를 가져오는 요인 중에는 발화 속도 (speaking rate)를 들 수 있겠는데 본 논문에서는 발화 속도에 따른 음성 인식의 성능 저하를 부분적으로 극복하는 방법에 대해 기술하였다.

발화 속도에 따른 인식율의 저하는 Mirghafori 등의 연구[2] 에서 빠른 발화 속도의 음성이 정상적인 발화 속도의 음성으로부터 학습된 모델을 사용하는 경우, 최대 4배의 인식 오류율을 나타내어 발화 속도가 음성 인식의 성능에 유의한 영향을 주는 요인임을 밝혔다. 이와 같은 발화 속도에 따른 오류율의 증가는 발화 속도가 빠른 음성의 특징 변수가 정상 속도의 음성에 비해 상이한 특징을 갖으며[11], 음소 생략 등에 의한 음소적 상이성 (phonological difference)이 고속 발화음 에서 유의하게 나타나는 것에 원인이 있다[2-3].

고속 발성음에 대해 인식율을 향상시키기 위한 방법은 크게 두 가지로 나눌 수 있다. 첫번째로 인식에 사용되는 모델을 고속 발성음에 적합하도록 적절히 변화시키는 것이고[2-3, 5, 7], 두번째는 모델은 고정시키고 입력된 음성을 변화시키는 것이다[8-9]. 전자의 대표적인 방법으로는, 음소 (phoneme)의 길이 모델 (duration model)을 빠른 음성의 길이 모델로 치환하는 기법[3, 6]이 있으며, 음성 인식의 통계 모델로 널리 사용되는 은닉 마코프 모델 (Hidden Markov Model; HMM)중 상태 천이 확률 (state transition probability)을 경험적인 방법으로 변경하는 방법이 있다[2]. 음성을 변형시키는 방법으로는 시간 영역의 음성 신호를 천천히 발성하는 소리로 시간축 변환 (time scale modification) 하여 특징 파라메터를 얻는 방법[8]과 MFCC (Mel Frequency Cepstrum Coefficients)와 같은 특징 변수를 보간 (interpolation)에 의해 시간적으로 확장하는 방법[9] 등이 있다. 본 논문에서는 MFCC를 시간적으로 확장시키는 방법을 적용하였다.

특징 변수를 시간적으로 확장하기 위해서는, 주어진 음성의 발성 속도에 따라 적응적으로 확장해야 한다. Richardson 등의 연구[9]에서는 음성 인식의 과정에서 얻어지는 음소 경계 정보를 이용하여, 각 음소의 길이를 구하고, 각 음소의 길이에 대한 확률 모델을 통해 음소별 확장 계수 (scaling factor)를 구하도록 하였다. 한 문장에 대한 확장 계수는 음소별 확장 계수의 평균으로 주어진다. 이와 같은 방법은 음소 지속 시간에 대한 통계 모델을 파라메터의 변환에 이용한 방법으로, 기존의 음소 지속 시간 모델을 변경하는 방법과 유사한 방법으로 볼 수 있다. 이 방법에는 몇 가지 단점이 있다. 첫 번째는 음소의 경계 정보를 추출하기 위한 부가적인 음성 인식이 수행되어야 한다는 점이고, 두 번째는 확장 요소의 결정 방법과 음성 인식의 성능 향상간에 직접적인 연관성이 없다는 점이다. 확장 계수는 단순히 음소의 지속 시간을 해당 음소에 대해 가장 자주 발생하는 지속 시간으로 변환시킬 뿐이며, 인식율이 최대화 되는 것을 보장하는 것은 아니다.

본 논문에서는 이와 같은 단점을 해결하기 위해 다음과 같은 방법이 제안되었다. 첫 번째로, 주어진 음성의 발화 속도는 비교적 간단한 계산만으로 구하도록 하였다. 제안 방법은 단위 시간당 음절 (syllable)의 개수가 발화 속도와 유의한 연관 관계를 갖는다는 가정[11] 에 바탕을 두어 MFCC의 계산 과정의 중간 단계에서 얻어지는 대역별 에너지값을 이용해 모음 구간을 판정하고, 이들 구간의 개수로 발화 속도를 추정하도록 하였다.

두 번째로 주어진 음성을 발화 속도에 따라 분류하고, 분류된 음성에 대해 개별적인 확장 계수를 결정하는 방법으로 최대 우도 (maximum likelihood) 기법이 적용되었다. 제안된 방법은 발화 속도의 분류를 위한 문턱치의 집합과 최적의 확장 계수를 반복적인 방법에 의해 추정하도록 하였으며, 점진적으로 우도를 증가시켜 궁극적으로는 최대 우도를 갖는 분류 규칙과 확장 계수들이 얻어지도록 하였다. 제안된 기법의 성능 평가를 위해 10자리 이동 전화 번호에 대해 음성 인식을 수행하였으며, 실험 결과를 제시하여 기존 방법간의 성능을 비교하였다

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 제안된 기법의 전체적인 블록도를 제시하고, 세부적으로 음성의 발화 속도 추정 방법, 최대 우도에 바탕을 둔 음성 분류 및 확장 방법을 설명하였다. 3장에서는 모의 실험 결과를 제시하였으며 4장의 결론에서 본 논문을 끝맺었다.

II. 고속 발화음에 대한 음성 인식 기법

제안된 음성 인식기의 블록도를 그림 1에 제시하였다. 입력된 음성은 먼저 인간의 귀와 유사한 특성을 갖는 대역 통과 필터들을 통과하고 각 대역의 신호들을 이용하여 발성 속도를 추정한다. 음성 인식의 특징 변수인 MFCC를 얻기 위해 대역 통과 필터의 에너지는 이산 여현 변환 (Discrete Cosine Transformation; DCT) 되고,

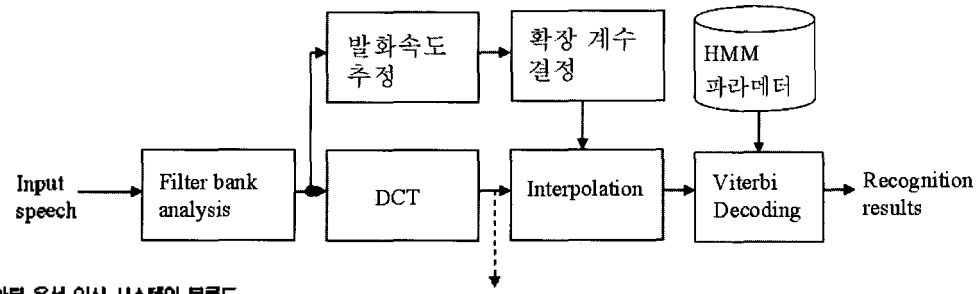

그림 1. 제안된 음성 인식 시스템의 블록도
Fig. 1. The block diagram of the proposed ASR system.

각 MFCC 계수는 시간축으로 보간에 의해 적절한 길이로 보정된다. 여기서 보간의 정도는 발성 속도에 따라 정해진 규칙으로 결정되는데 이 규칙은 최대 우도에 바탕을 둔 최적화 방법에 의해 생성된다. 각 블록에 대한 세부적인 설명은 다음과 같다.

2.1. 발화 속도의 추정

주어진 음성 신호로부터 발화 속도 (Rate Of Speech; ROS)를 추정하는 방법은 음성 인식과 유사한 방법을 이용하여 음성 신호를 음소 정보로 변환하여 추정하는 방법[2, 4, 7] [8-9]과 주어진 음성 신호에서 직접적으로 추정하는 방법[6, 11]이 있다. 전자의 방법은 음소 정보의 추출에 필요한 통계적 모델이 필요하며, 이를 위해 별도의 학습 데이터 및 학습 과정이 필요하다. 반면 후자의 방법은 이러한 모델이 필요치 않으며, 비터비 (Viterbi) 디코딩과 같은 인식 과정이 불필요하므로 계산량이 적다는 장점이 있다. 음성 인식에 포함되는 발화 속도의 추정 방법은 시스템 전체로 볼 때, 전처리 과정 (preprocessing)에 포함되므로 적은 계산량이 바람직하다. 이러한 이유로 본 논문에서는 음성 신호의 파형으로부터 직접적으로 발화 속도를 추정하는 방법을 적용하였다.

제안된 발화 속도의 추정 방법은 단위 시간 내에 포함되는 음절의 개수가 발성 속도와 유의한 연관 관계를 갖으며, 음성중 모음에 해당하는 구간은 특정 대역에 높은 에너지가 분포할 것이라는 두 가정을 바탕으로 하고 있다. 실제로 Pfau등 [11]은 특정 대역의 에너지를 나타내는 변형 라우드네스 (modified loudness)를 도입하고, 이를 이용해 음성의 모음 구간을 추정하고, 단위 시간당 모음의 개수를 구함으로서 발성 속도를 추정하는 방법을 제안하였다. 본 논문에서도 이와 유사한 기법이 제안되었는데, Pfau가 제안한 modified loudness의 정의는 다음과 같다[11].

$$D(t) = N_u(t) - N_o(t) = \sum_{\nu=3}^{15} N_\nu(t) - \sum_{\nu=20}^{22} N_\nu(t)$$
$$N_m(t) = Max\{D(t), 0\} \quad (1)$$

여기서 $N_\nu(t)$는 임계 대역 (critical band) ν 내의 에너지를 나타낸다. 즉, 식 (1)로부터, modified loudness 함수는 특정 임계 대역내의 에너지 합과 차로 주어짐을 알 수 있다.

모음 구간의 판정은 먼저 $N_m(t)$를 5-pont 중간값 필터 (median filter)를 통해 스무딩 시키고 유의한 극대값을 검출함으로서 구현된다. 발화속도는 아래의 식으로 나타낼 수 있다.

$$ROS = \frac{N_{peak}}{L} \quad (2)$$

여기서 N_{peak}는 스무딩된 $N_m(t)$의 피크 개수를 나타내며 L은 음성의 전체 길이를 나타낸다. 모음의 특성은 언어 마다 다른 특성을 나타내며, Pfau 의 연구는 영어권 언어를 대상으로 하였기 모음 판정을 위한 modified loudness 값 또한 한국어에 맞게 적절히 변경되어야 한다. 본 논문에서는 수 차례의 실험을 통해 (1)의 $N_u(t)$와 $N_o(t)$ 각각에 대해 임계 대역 구간을 결정하였는데, 결론적으로 $N_u(t)$는 임계 대역 2-8 바크 (bark) 에 해당하는 에너지의 합을, 그리고 $N_o(t) = 0$으로 설정하는 경우 최적의 성능을 나타내었다.

2.2. ROS 분류 및 확장 계수의 결정

주어진 음성에서 ROS를 추정하였으면, 이에 따라 적절히 MFCC 를 시간축으로 확장해야 한다. 이를 위해 본 논문에서는 ROS에 따라 음성 신호를 분류하고, 분류된 각각의 음성 신호에 대해 MFCC의 확장 계수를 결정하도록 하였다. 이 때의 분류는 그림 2와 같이 크기 순으로

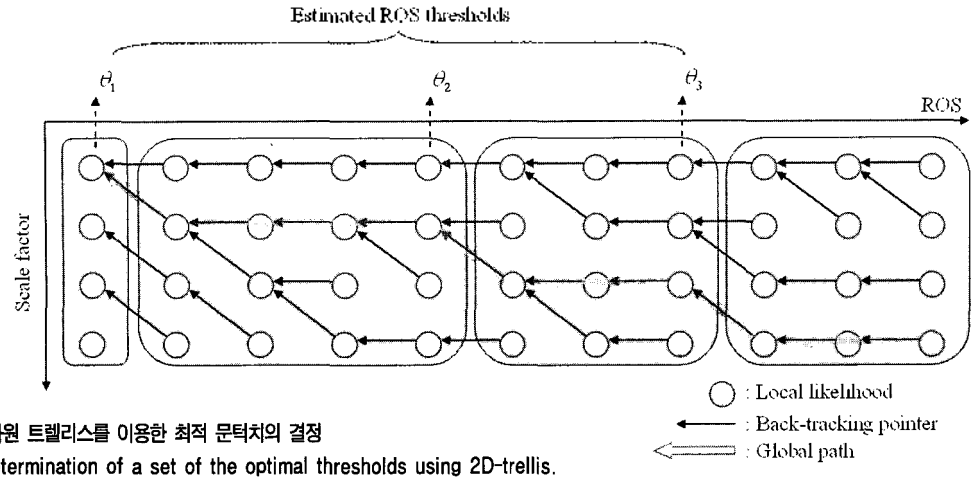

그림 3. 2차원 트렐리스를 이용한 최적 문턱치의 결정
Fig. 3. Determination of a set of the optimal thresholds using 2D-trellis.

정렬된 일련의 문턱치 (threshold)와 ROS 값을 비교하여 선택된 구역에 대응하는 확장 계수를 MFCC의 확장에 사용하게 된다.

본 논문에서는 일련의 문턱치 집합 $\Theta=\{\theta_0,\theta_1,...\theta_M\}$ 과 각 구역에 대응하는 확장 계수의 집합 $\Gamma=\{\gamma_1,...\gamma_M\}$ 을 구하기 위해 최대 우도 (Maximum Likelihood; ML) 기법이 사용되었다. ML 기법을 적용하기 위해 먼저 아래와 같은 입력 특징 벡터 열 $\mathbf{X}$에 대한 로그 우도 율 (log likelihood ratio) 을 정의하였다.

$$L(\mathbf{X},\Lambda,\Theta,\Gamma)=\sum_{m=1}^{M}\sum_{\mathbf{x}\in S_m}\log[p(f(\mathbf{x},\gamma_m)|\Lambda)] \tag{3}$$

여기서 Λ는 음성 인식에 사용되는 모델 파라메터, 즉 HMM 과 관련된 여러 파라메터 집합을, M 은 분류 개수를 나타내며 문턱치 집합내의 각 문턱치는 $\theta_m \le \theta_{m+1}$, $\theta_0=-\infty$, $\theta_M=\infty$ 을 만족한다. S_m 은 다음과 같다.

$$S_m=\{\mathbf{x}|\theta_{m-1}<ROS(\mathbf{x})\le\theta_m\} \tag{4}$$

여기서 $\mathbf{x}=\{\vec{x}_1,\vec{x}_2,...,\vec{x}_N\}$는 하나의 문장에 포함되는 모든 MFCC 벡터들을 나타내며, 따라서 S_m은 ROS가 θ_{m-1} 보다 크고 θ_m 보다 작거나 같은 모든 문장에 포함되는

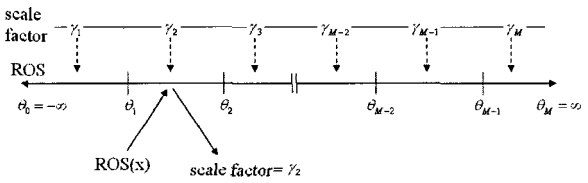

그림 2. ROS를 이용한 확장 계수의 선택
Fig. 2. Selection of the scale factor using ROS.

MFCC 벡터의 집합을 나타낸다. $f(\mathbf{x},\gamma_m)$는 확장 계수 γ_m 에 의해 보간된 MFCC 벡터를 나타낸다.

본 논문에서 정의한 최적화는 (3)으로 주어지는 우도를 최대화 하는 문턱치의 집합 Θ과 확장 계수의 집합 Γ를 구하는 문제로 나타낼 수 있다. 즉,

$$\Theta^*,\Gamma^*=\arg\max_{\Theta,\Gamma} L(\mathbf{X},\Lambda,\Theta,\Gamma) \tag{5}$$

이와 같은 문제를 해결하기 위해, 본 논문에서 반복 추정 방법 (iterative method) 를 사용하였다. 이 방법은 초기 문턱치의 집합 Θ를 이용하여 $L(\mathbf{X},\Lambda,\Theta,\Gamma)$를 최대화하는 확장 계수 집합 $\tilde{\Gamma}$를 구하고, $\tilde{\Gamma}$를 이용하여 $L(\mathbf{X},\Lambda,\Theta,\tilde{\Gamma})$을 최대화하는 문턱치 집합 $\tilde{\Theta}$을 반복적으로 구하는 것이다.

이러한 과정을 통해 $L(\mathbf{X},\Lambda,\Theta,\Gamma)$이 지속적으로 증가될 수 있으며, 궁극적으로 식 (5)를 만족하는 최적의 Θ^*,Γ^*에 근접한 값을 구할 수 있게 된다. 반복 추정 기법의 각 과정을 자세히 살펴보면 다음과 같다.

과정0) 초기화: 학습에 필요한 MFCC 데이터 $\mathbf{X}$를 준비하고, 인식에 필요한 HMM 파라메터 Λ를 Baum-Welch 방법[1] 등을 이용하여 구성한다. 초기 확장 계수 집합 $\Gamma^{(0)}=\{\gamma_1^{(0)},\gamma_2^{(0)},...,\gamma_M^{(0)}\}$을 적절한 방법으로 구성하고, $i=0$, $\lambda^{(0)}=-\infty$로 설정한다.

과정1) 최적 문턱치 집합의 추정: 주어진 $_{(i)}$를 이용하여 $L(\mathbf{X},\Lambda,\Theta,\Gamma^{(i)})$를 최대화하는 $\Theta^{(i)}$를 구한다. 본 논문에서는 이를 구하는 방법으로, 그림 3에 제시한 바와

같이 각각 오름차순으로 정렬된 ROS 값과 확장 계수들로 2차원 trellis를 구성하고, 여기에 비터비 (Viterbi) 알고리즘을 적용하여 최대 우도 관점에서 최적의 경로 (optimal path)를 찾도록 하였다. 최적의 경로를 통해, 동일한 확장 계수를 갖는 ROS 값의 범위를 알 수 있으며, 이를 통해 ROS의 경계값을 알 수 있다. 상세한 과정은 다음과 같다.

과정1-1) 학습 데이터에 포함된 모든 문장에 대해 ROS를 구하고, 오름차 순으로 정렬한다.
즉, $ROS_1 < ROS_2 < ... < ROS_N$. 여기서 N은 ROS 값의 총 개수이다. 다음으로 아래 식으로 정의되는 지역 우도값 (local likelihood) $l(n,m)$을 계산한다.

$$l(n,m) = \sum_{\mathbf{x} \in S_n} \log[p(f(\mathbf{x}, \gamma_m)], \ 1 \le n \le N, \ 1 \le m \le M \tag{6}$$

여기서 $S_n = \{\mathbf{x} | ROS(\mathbf{x}) = ROS_n\}$, 즉 ROS 값이 ROS_n 인 모든 문장에 포함되는 MFCC 벡터의 집합을 나타낸다. 따라서 즉, $l(n,m)$은 ROS 값이 n-번째 ROS 에 해당되는 문장을 m-번째 확장 계수 γ_m으로 시간축 확장하는 경우 로그 우도의 총합을 나타낸다. 확장 계수 γ_m도 ROS와 마찬가지로 오름차순으로 정리되었다 가정한다. 즉, $\gamma_1 < \gamma_2 < ... < \gamma_M$.

과정1-2) 위의 과정을 통해 구한 지역 우도값 $l(n,m)$을 이용하여 각 지점 (n,m)에서의 지역 최적 경로 (locally optimum path)를 구한다 (7)

$$L(n,m) = \max_{i=m, m-1} \{L(n-1,i)\} + l(n,m)$$
$$\Psi(n,m) = \arg\max_{i=m, m-1} \{L(n-1,i)\} \quad , \ 1 \le n \le N, \ 1 \le m \le M$$

여기서 $L(n,m)$은 m-번째 확장 계수에 대한 n-번째 ROS 까지의 누적 우도값을 나타낸다. $\Psi(n,m)$은 해당 지점에서의 역트랙 포인터 (back-tracking pointer)를 나타내는데, 최종적으로 추정된 문턱치들이 오름차순으로 구성되도록 다음과 같은 제한을 가하였다. (그림 3 참고)

$$\Psi(n,m) = \begin{cases} 1 & , \text{ if } \mathrm{m} = 1 \\ m-1 & , \text{ if } \mathrm{n} = \mathrm{m} \end{cases} \tag{8}$$

과정1-3) 역트랙 포인터를 이용하여 최적의 경로를 구한다. 위의 과정과 마찬가지로 문턱치가 오름차순으로 구성되도록 끝점에서 아래와 같은 제한을 가하였다.

$$\varphi(n) = \begin{cases} M & , \text{ if } n = N \\ \Psi(n+1, \varphi(n-1)) & , \text{otherwise} \end{cases} \tag{9}$$

최종적으로, ROS의 문턱치는 그림 3에 나타난 바와 같이, 경로 $\varphi(n)$의 방향이 전환되는 부분에서 결정된다. 즉,

$$\Theta^{(i)} = \{\theta^{(i)} = ROS_{n-1} \,|\, \varphi(n-1) \neq \varphi(n), 1 \le n \le N\} \tag{10}$$

과정2) 최적 확장 계수 집합의 추정: 과정1) 에서 구한 $\Theta^{(i)} = \{\theta_0^{(i)}, \theta_1^{(i)}, ..., \theta_M^{(i)}\}$를 이용하여 $L(\mathbf{X}, \Lambda, \Theta^{(i)}, \Gamma^{(i)})$를 최대화하는 $\Gamma^{(i+1)}$를 구한다. 즉, ROS가 m-번째 구간에 포함되는 모든 문장에 대해 우도를 최대화 하는 확장 계수를 구한다.

$$\gamma_m^{(i+1)} = \arg\max_{\gamma} \{\sum_{\mathbf{x} \in S_m} [\log p(f(x, \gamma) \,|\, \Lambda)]\}, \ 1 \le m \le M \tag{11}$$

여기서 S_m은 식 (4)에 제시한 것과 동일하다. 확장 계수 γ와 우도간의 관계는 analytical 하게 나타낼 수 없으므로 (11)을 만족하는 확장 계수 $\gamma_m^{(i+1)}$는 수학적으로 얻어질 수 없다. 본 논문에서는 확장 계수의 미세한 변화에 우도값이 민감하게 반응하지 않는다는 가정에 따라 확장 계수값을 몇 개로 제한하고, 이들 개별 계수에 대해 (11)의 우도 총합을 각각 구하고, 이들 중 가장 큰 값을 나타내는 확장 계수값을 최적 확장 계수로 선택하였다. i-번째 반복과정에서의 최종적인 확장 계수 집합 $\Gamma^{(i+1)}$은 (11)을 만족하는 개별 확장 계수들로 구성된다.

$$\Gamma^{(i+1)} = \{\gamma_1^{(i+1)}, \gamma_2^{(i+1)}, ..., \gamma_M^{(i+1)}\} \tag{12}$$

과정3) 수렴 여부 조사: i-번째 반복과정에서 구한 $\Theta^{(i)}$와 $\Gamma^{(i+1)}$를 이용하여 (3)으로 주어지는 우도값 $\lambda^{(i)} = L(\mathbf{X}, \Lambda, \Theta^{(i)}, \Gamma^{(i+1)})$을 구한다. 우도값이 이전의 반복과정과 비교하여 거의 변화가 없으면 반복과정을 중단하고

현재 추정된 $\Theta^{(i)}$ 와 $\Gamma^{(i+1)}$ 을 최종적인 결과로 간주하고 그렇지 않다면 $i=i+1$ 하고 과정1)~과정3)을 반복한다.

2.3. MFCC 벡터의 시간축 확장

MFCC 벡터의 시간축 확장은, MFCC 벡터를 구성하는 각 성분에 대해 소수점 표본화 (fractional sampling)를 수행하고 이를 다시 본래의 샘플 간격으로 재구성 함으로서 이루어진다. Richardson의 연구[9] 에서는 대역제한 샘플링 함수 (sinc-function)을 이용하는 방법과 단순히 MFCC 벡터를 복사하는 방법 등이 사용되었는데, 두 방법간의 유의한 차이는 없는 것으로 보고하였다. 본 논문에서도 쌍선형 보간 (bi-linear interpolation), 대역제한 샘플링 함수를 이용한 보간, 복사 방법의 3가지 방법을 적용하였다. 실험 결과에 따르면 Richardson의 방법과 달리, MFCC를 단순히 복사하는 방법이 나머지 두 방법에 비해 다소 저하된 성능을 나타내었다. 이러한 결과에 따라, 본 논문에서 보간 필터에 의해 시간축 확장을 구현하였으며, 계산량이 상대적으로 적은 쌍선형 보간 방법을 사용하였다.

III. 실험 및 결과

본 논문에서 제안된 기법의 성능 평가를 위해 실제 고속의 발성음에 대해 음성 인식을 수행하고 결과를 살펴보았다. 음성 인식의 대상이 되는 문장은 이동 통신 전화번호인 10자리 숫자음을 사용하였으며, 각 숫자음은 트라이폰 부모델 (triphone subword model)이 직렬 연결된 형태로 표현하였다. 사용된 부모델의 개수는 총 24개인데, 이중 2개의 부모델은 통계적인 방법에 따라 공유 모델 (shared-model)이 사용되어 실제적으로는 22개의 HMM이 사용되었다. 음성 파라메터는 13차 MFCC가 이용되었으며, 차분 (delta) 값 및 가속 (acceleration) 값을 포함하는 총 39개의 변수가 HMM의 생성 과 음성 인식에 사용되었다. 이들 변수의 계산 및 학습 조건은 표 1에 제시 하였다.

음성은 남, 녀 각각 8명으로 구성된16명의 화자에 의해 10종류의 전화번호를 10번 반복 발성하여 녹음하였다. 또한 빠르게 발성하는 음성에 대한 성능을 평가하기 위해 동일한 화자들로부터 동일한 전화번호를 10회 빠르게 발성한 음성을 추가적으로 취득하였다. HMM의 학습

표 1. 실험 조건
Table 1. Experiment condition.

MFCC 차수	13 (0차 계수 포함)
분석 프레임 이동 길이	10 msec
분석 프레임 길이	25 msec
HMM 종류	연속 모델
HMM 형태	좌-우 모델
HMM 상태 (state) 수	5 (묵음은 3)
혼합 가우시안의 개수	5

표 2. 각 음성 데이터에 대한 인식율 과 평균ROS 와의 관계
Table 2. Recognition rates and average ROS for each speech corpus.

음성 데이터	평균 ROS (vowels/sec)	인식율 (%)
학습 데이터	8.21	96.59
테스트 데이터 (정상속도)	10.37	92.26
테스트 데이터 (고속)	14.42	83.32

에는 8명의 화자를 임의로 선택하여 발성 속도에 관계없이 10종류 전화 번호에 대해 20 번 반복한 음성이 사용되었는데, 이는 총 1600개의 문장, 19200개의 단어에 해당한다. 테스트 데이터의 구성에는 학습 시 사용되지 않는 나머지 8명의 화자 음성이 사용되었는데, 정상 속도로 발성한 800개의 문장들과 고속으로 발성한 800개의 문장들을 각각 정상 속도 테스트 데이터, 고속 테스트 데이터로 구분하여 구성하였다.

본 논문에서는 또한 MFCC의 시간축 확장을 적용한 음성 인식 기법과 기존 방법간의 성능 비교를 위해, 고속의 발성 음성으로 학습한 별도의 HMM을 생성하여 음성 인식 성능의 비교 대상으로 삼았다.

본 논문에서는 설명의 편의를 위해, 1600개 문장/19200개의 단어를 포함하는 학습 데이터로 구성된 HMM을 HMM-A라 칭하고, 학습 데이터를 ROS값에 따라 2개로 분할하여, 고속의 발성음으로 판정된 음성들로 학습된 HMM을 HMM-B로, 그리고 나머지 발성음으로 학습된 HMM을 HMM-C라 칭하였다.

3.1. 각 테스트 데이터에 대한 음성 인식 성능 비교

표 2에 학습 데이터, 정상 속도로 발성한 테스트 데이터, 빠른 속도로 발성한 테스트 데이터에 대한 발화 속도 (ROS) 와 단어 인식율이 나타나있다. 학습 데이터와 정상 속도의 테스트 데이터 간 인식율은 각각 96.59% 와 92.26% 로서, 두 데이터간의 유의한 차이는 관찰되

지 않았다. 두 데이터군의 평균 발화 속도는 각각 8.21과 10.37로서 테스트 데이터가 다소 빠르게 발성한 음성임을 알 수 있다. 발화 속도의 차이는 학습 데이터와, 테스트 데이터에 각기 다른 화자들을 선택하였고, 이들 화자들의 발성 스타일이 서로 다른 것에 기인된 것으로 판단된다. 한편 빠른 발성음에 대한 인식율은 83.32%로 정상 속도의 발성음에 비해 약 9% 낮은 인식율을 나타내었다. 발화 속도면 에서는 10.37 대 14.42 로 빠른 발성음이 25% 정도 높은 ROS 값을 나타내었다.

3.2. 제안된 기법을 적용한 음성 인식 시스템의 성능

제안된 음성 인식 기법의 고속 발성음에 대한 유용성을 알아보기 위해, 고속 발성음의 MFCC를 시간적으로 확장하고, 인식율을 구하였다. 성능 비교의 대상으로, 시간적 확장이 고려되지 않는 본래의 MFCC로 인식을 수행하는 경우와, 발성 속도에 따라 별개로 구성된 2개 HMM을 이용하는 경우를 고려하였다.

본 실험에서는 2.1절에서 제시한 바와 같이 몇 개의 제한된 시간축 확장 계수만을 사용하였는데, 2.0 이상의 시간축 확장에서는 MFCC의 특성 변화에 따른 인식율의 저하가 유의하게 관찰되었다. 따라서 실험에서는 시간축 확장 계수로 1.05에서 1.95 까지 0.05 단계 (step-size)로 증가시킨 값을 사용하였다 (총 19종류). 여기서 Step-size를 감소시키면 보다 정확한 시간축 확장 계수를 얻는데 유리하나, 제안된 알고리즘이 가능한 모든 확장 계수에 대해 전 탐색 (full search)을 수행하므로 추정 시간이 오래 걸린다는 문제가 있다. 경험적으로, 0.5 미만의 step-size를 사용하는 경우, 성능이 크게 향상되지는 않았다.

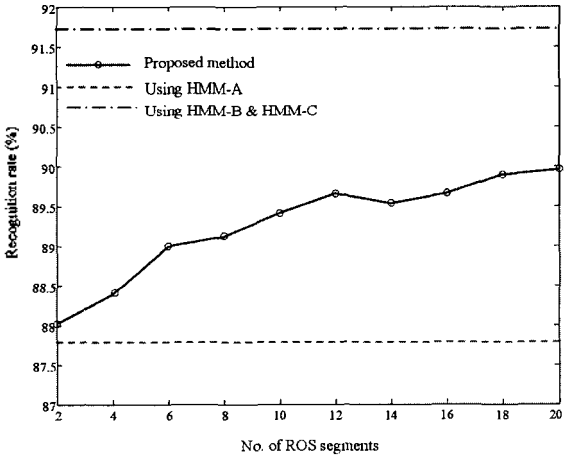

그림 4. ROS의 분류 개수에 따른 인식율
Fig. 4. Recognition rate according to the number of ROS segments.

그림 4에 ROS를 분류하는 개수, 즉 실제적으로 사용되는 확장 계수의 개수에 따른 인식율이 제시되었다. 이 결과는 고속 발성음과 정상 발성음이 모두 포함된 학습 데이터에 대한 인식 결과이다. 그림에서는 성능 비교를 위해 발화 속도에 따라 2개의 독립된 HMM을 사용하는 경우와 1개의 HMM을 사용하는 경우의 인식율이 함께 나타나있다.

그림에서 보면, 사용된 확장 계수의 개수에 따라 인식율도 완만하게 상승하는 것을 알 수 있으며, 20개의 확장 계수, 즉 ROS 를 20개로 분할하여 별개의 시간축 확장을 수행하는 경우 최대 인식율 89.95%가 얻어짐을 알 수 있다. 이 값은 표 2에 제시한 두 종류의 테스트 데이터에 대한 전체 인식율 87.77% 보다 2.18% 증가된 값이며, 오차율 면에서는 12.23% 대 10.05%로 약 17.8%의 오차 감소가 이루어졌음을 의미한다.

또한 그림에는 제시되어 있지 않지만, Richardson이 제안한 음소 지속 시간을 이용한 MFCC 확장 방법[9]과 비교하여 ROS 분류 개수를 2개로 설정한 경우 약 0.05%의 향상된 인식율을 얻을 수 있었으며 (87.97% 대 87.92%), ROS 분류 개수를 증가시킴에 따라 성능 차이가 더욱 심하게 나타났다. 이는 ROS를 단지 2개로 구분하여 시간축 확장을 수행하더라도, 음소 지속 시간 정보만을 이용한 시간축 확장 기법 보다 우수한 성능을 나타냄을 의미한다. 이는 분류 와 확장 계수의 설정이 경험적으로 이루어지는 Richardson 방법에 비해, 제안된 기법은 우도를 최대화하는 관점에서 최적화된 분류, 확장 계수가 결정되기 때문이다.

그러나 그림 4에서 두 개의 HMM을 사용한 경우 인식율은 91.74%로, 최대의 ROS 분류 개수 (=20)를 사용한 경우와 비교하여 보다 높은 값을 나타내었다. 실험 결과에 따르면, 분류 개수를 20개 이상으로 설정하는 경우 과추정 (overestimation) 문제로 인한 학습 데이터와 테스트 데이터간의 성능 불균형이 심하게 나타났다. 따라서 시간축 확장을 통해 얻을 수 있는 최대 인식율을 89.95% 라 한다면 (ROS의 분류 개수=20인 경우), 두 개의 HMM 을 사용하는 경우와 1.79% 의 인식율 차이를 나타냄을 알 수 있다. 이러한 인식율 저하는 시간축 확장을 통해 변형된 MFCC가 단순히 시간적으로 확장된 신호만을 의미할 뿐 이며 정상 속도의 발성음과 고속 발성음 간의 음향적인 차이 (acoustical difference)가 보상되지는 못했기 때문이다. 실제로 고속 발성음과 정상 발성음의 비교에 관한 최근의 연구[10] 을 보면, 고속 발

성음은 조음화 (co-articulation) 현상이 유의하게 나타나며, 이로 인해 포만트 (formant) 주파수가 변화될 수 있다고 보고되었다. 따라서 보다 높은 인식율을 얻기 위해서는, 단순히 선형 시간축 확장이 아닌, MFCC 의 특성 자체를 정상 속도의 발성음과 유사하도록 변경 시키는 방법이 연구되어야 할 것이다.

VI. 결 론

본 논문에서는 대화체 음성에 대한 인식율을 향상시키는 방편의 하나로, 고속으로 발성한 음성에 대해 강인한 특성을 갖는 음성 인식의 방법을 제안하고 성능을 평가하였다. 제안된 방법은 음성 인식의 특징 변수를 단순히 시간축으로 확장하는 방법만으로 어느 정도의 성능 향상이 이루어질 수 있는가를 고찰하였으며, 음성의 빠르기에 대한 확장의 정도는 최대 우도 방법에 근거한 최적화 방법을 통해 결정하였다.

10자리 아동 전화 번호를 이용한 음성 인식의 실험 결과, 제안된 기법은 단지 1개의 HMM을 사용하는 기법에 비해 향상된 음성 인식율을 얻었으며, 비교적 적은 계산량으로 부가적인 HMM 없이 성능을 향상시킬 수 있는 방법임을 입증하였다. 이는 제안된 기법이 부가적 HMM 파라메터를 저장하기 위한 메모리 공간 대신, 상대적으로 적은 공간을 차지할 것으로 예상되는 문턱치의 값과 각 구획별 확장 계수들을 사용함으로서 음성 인식 시스템 구성에 필요한 전체 메모리 공간을 절약할 수 있을 것으로 기대된다.

본 논문에서 제안한 최적의 ROS 분할 방법은 임의의 스칼라 변수를 분할하여 각 분할된 영역에 대해 최적의 파라메터를 구하는 문제의 해결 방법으로서, 유사한 문제에 적용될 수 있을 것으로 판단된다. 예로서, 제안된 기법을 이용하여 음성 신호가 가지고 있는 여러 운율 정보들 (피치, 에너지 등)에 따라 최적의 변환 기법을 설계할 수 있을 것으로 사료된다.

참 고 문 헌

1. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, **77**, Issue 2, 257-286, 1989.
2. N. Mirghafori, E. Fosler and N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes," The proceedings of EUROSPEECH95, 491-494, Madrid, Spain, September 1995.
3. N. Mirghafori, E. Fosler and N. Morgan, "Towards robustness to fast speech in ASR," The proceedings of ICASSP96, 335-338, Atlanta, USA, 1996.
4. M. J. Russell, K. M. Ponting and M. J. Tomlinson, "Measure of local speaking-rate for automatic speech recognition," IEE Electronics Letters, **35** (10), 787-789, 1999.
5. M. H. Nguyen and G. W. Cottrell, "A technique for adapting to speech rate," *The proceedings of the 1993 IEEE-SP* workshop, 6-9, 382-391, September 1993.
6. R. Fallthauser, T. Pfau and G. Ruske, "On-line speaking rate estimation using Gaussian mixture models," The proceedings of ICASSP2000, 1355-1358, 2000.
7. J. Zheng, H. Franco and A. Stolcke, "Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition," Speech Communication, **41**, 273-285, 2003.
8. 이기승, "시간축 변환을 이용한 음성 인식기의 성능 향상에 관한 연구," 한국음향학회지, **23** (6), 462-472, 2004년 8월.
9. M. Richardson, M. Hwang, A. Acero and X. Huang, "Improvements on speech recognition for fast talkers," The proceedings of EUROSPEECH1999, 411-414, 1999.
10. L. Deng, D. Yu, and A. Acero, "A quantitative model for formant dynamics and contextually assimilated reduction in fluent speech," The Proceedings of the ICSLP, Oct.4-8, 2004, Jeju Island, Korea, No. WeA501 **20**, 501-504.
11. T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," The Proceedings of the ICASSP 98, 945-948, 1998.

저자 약력

• 이 기 승 (Ki-Seung Lee)

1991년 2월: 연세대학교 전자공학과 (공학사)
1993년 2월: 연세대학교 대학원 전자공학과 (공학석사)
1997년 2월: 연세대학교 대학원 전자공학과 (공학박사)
1997년 3월~1997년 9월: 연세대학교 신호처리 연구센터 선임연구원
1997년 10월~2000년 9월: AT&T Shannon Lab 연구원
2000년 11월~2001년 8월: 삼성종합기술원 HCI Lab 전문연구원
2001년 9월~현재: 건국대학교 정보통신대학 전자공학부 조교수
*주관심분야: 음성 합성, 운율제어, 음성변환,음성 부호화기 등