한국음향학회지 제23권 제6호 pp. 462~472 (2004)

시간축 변환을 이용한 음성 인식기의 성능 향상에 관한 연구

Study on the Improvement of Speech Recognizer by Using Time Scale Modification

이 기 승*
(Ki-Seung Lee*)

*건국대학교 정보 통신 대학 전자 공학과

(접수일자: 2003년 11월 19일; 수정일자: 2004년 7월 19일; 채택일자: 2004년 8월 24일)

본 논문에서는 자동 음성 인식기의 성능 저하를 일으키는 요인으로서, 발성 속도의 변동에 따른 성능 저하를 보상하기 위한 기법을 제안하였다. 새로운 기법의 제안에 앞서서, 먼저 발성 속도의 변화에 따른 기존의 은닉 마코프 모델을 이용한 음성 인식기의 성능을 정량적으로 분석하였다. 이러한 분석을 통해 발성 속도에 따른 유의한 성능 저하를 관찰하고, 주어진 음성으로부터 발성 속도를 정량적으로 나타낼 수 있는 변수를 도입하였다. 발성 속도를 학습 시 사용한 음성과 유사하게 변화시키기 위해 본 논문에서는 음성 신호에 대한 시간축 변환을 사용하였으며, 최종적으로 발성 속도에 따라 선택적으로 시간축 변환을 적용하여 발성 속도의 변동에 따른 음성 인식의 성능 저하를 보상할 수 있는 기법을 제안하였다. 10자리의 이동통신용 전화번호를 이용한 음성 인식의 실험을 통해, 제안된 기법은 빠르게 발성하는 음성에 대해 15.5%의 오류율 감소를 가져오는 것을 확인할 수 있었다.

핵심용어: 자동 음성 인식, 시간축 변환, 발성 속도
투고분야: 음성처리 분야 (2.5)

In this paper, a method for compensating for the performance degradation of automatic speech recognition (ASR) is proposed, which is mainly caused by speaking rate variation. Before the new method is proposed, quantitative analysis of the performance of an HMM-based ASR system according to speaking rate is first performed. From this analysis, significant performance degradation was often observed in the rapidly speaking speech signals. A quantitative measure is then introduced, which is able to represent speaking rate. Time scale modification (TSM) is employed to compensate the speaking rate difference between input speech signals and training speech signals. Finally, a method for compensating the performance degradation caused by speaking rate variation is proposed, in which TSM is selectively employed according to speaking rate. By the results from the ASR experiments devised for the 10-digits mobile phone number, it is confirmed that the error rate was reduced by 15.5% when the proposed method is applied to the high speaking rate speech signals.

Keywords: Automatic speech recognition, Time scale modification, Speaking rate
ASK subject classification: Speech signal processing (2.5)

I. 서 론

음성 신호 처리와 관련된 신호 처리 기법의 발달에 따라 입력된 음성을 자동적으로 인식하는 자동 음성 인식기 (Automatic Speech Recognizer; ASR)[1-3] 개발이 가능하게 되었다. 자동 음성 인식기는 기본적으로, 학습 (training) 에 의해 인식하고자 하는 단어들의 기본 패턴 (prototype pattern) 을 작성하고, 온라인 인식 과정에서는 입력된 음성의 특징 변수를 기본 패턴에 포함된 변수들과 비교하여 가장 유사한 패턴을 인식된 결과로 출력하게 된다.

은닉 마코프 모델 (Hidden Markov Model; HMM) 이

책임저자: 이 기 승 (kseung@kkucc.konkuk.ac.kr)
143-701 서울시 광진구 화양동 1번지
건국대학교 정보통신대학 전자공학과 1417호
(전화: 02-450-3489; 팩스: 02-3437-5235)

음성 인식에 성공적으로 적용된 이후, 음성 인식의 분야는 새로운 알고리즘의 개발과 이를 탑재할 수 있는 하드웨어의 성능 향상에 따라 많은 성능 향상이 이루어졌다. 그러나 아직까지 음성 인식은 제한적인 용도로 이용되고 있는 것이 사실인데, 이는 사용자가 불편을 느낄 수 없는 성능을 얻을 수 없다는 점과 사용 환경에 따라 성능 편차가 다소 크게 나타난다는 점에 그 원인을 찾을 수 있다.

음성 인식의 성능 저하를 일으키는 주된 이유로, 학습 시에 사용된 음성 신호와 온라인 인식시에 입력되는 음성 신호의 차이를 들 수 있겠다. 즉, 학습 시에 사용된 음성의 녹음 환경, 잡음 정도, 화자의 발성 스타일 등이 온라인 음성 인식 시에 입력되는 음성과는 서로 다른 특징을 갖을 수 있으며, 이는 학습 시에 작성된 기준 패턴과의 유사성을 떨어뜨려 잘못된 인식 결과를 초래할 수 있다. 본 논문에서는 이와 같은 학습 시와 온라인 인식시의 음성 신호 차이를 발화 속도 (speaking rate) 면에서 분석하고, 이러한 차이를 보상하여 결과적으로 발화 속도의 변동에 강인한 음성 인식 기법을 제안하고자 한다.

발화 속도에 따른 음성 인식의 성능 저하는 Mirghafori 등의 연구[1]에서 빠른 발화 속도의 음성이 정상적인 발화 속도의 음성을 동일한 HMM을 사용하여 인식하는 경우, 최대 4배의 인식 오류율 (recognition error rate)을 나타내어 발화 속도가 음성 인식의 성능에 영향을 주는 주된 요인임을 입증하였다. 이와 같은 발화 속도에 따른 오류율의 증가는 발화 속도가 빠른 음성으로부터 추정된 특징 변수가 정상 속도의 음성에 비해 차이를 나타내며, 음소의 생략 등으로 인하여 음소적 상이성 (phonological difference)이 빠른 발성음 에서 유의하게 나타나는 것에 원인이 있다[1,2].

특히, 낭독체 음성 (reading speech)과 비교하여 대화체 음성 (conversational speech)은 동일한 화자일지라도 발화 속도의 차이가 심하게 나타나며[3], 이는 대화체 음성의 인식 성능 저하를 일으키는 한 요인이 발화 속도에 있음을 의미한다. 따라서 실생활에 보다 널리 사용되는 대화체 음성을 대상으로 하는 음성 인식 시스템은 발화 속도를 고려해야 인식율의 향상을 가져올 수 있다.

빠른 속도의 발화 음성에 대처하는 가장 간단한 방법은, 빠른 발화 음성들로부터 별개의 기준 패턴을 만들고, 이를 온라인 인식과정에 사용하는 것이다. 예로서, 학습 데이터의 생성 시 화자들에게 특별히 빠른 발성을 요구하여 별개의 데이터베이스를 구성하여 이로부터 HMM를 생성하고, 인식 과정에서는 정상 속도의 HMM과 빠른 속도의 HMM을 동시에 이용하는 방법을 들 수 있다. 이러한 방법은 특별히 빠르게 발성한 음성들로 추가적인 학습 데이터를 구성해야 하며, 인식 시 두 배로 증가된 기준 패턴을 고려해야 하므로 필요한 저장 공간과 계산량이 증가된다는 단점이 있다.

이와 같은 단점을 해소하기 위한 방안으로, Morgan 등은 정상 속도의 음성에서 생성된 HMM을 변형시키는 방법을 제안하였다[2]. 이 방법은 고속의 발성음 에서 얻어진 특징 변수가 시간적으로 빠르게 변화한다는 사실에 착안하여, HMM을 구성하는 변수의 하나인 상태 천이 확률 (state transition probability) 값을 경험적인 방법으로 재조정하도록 하였다. 즉, 동일한 상태에 지속될 확률값을 상대적으로 줄이고, 상태가 바뀔 확률값을 증가시킴으로서, 빠르게 변동되는 음성의 특성을 반영하도록 하였다. 이와 같은 방법은 고속의 발화 음성에 대한 별도의 HMM을 필요치 않으며 인식율도 별도의 HMM을 사용한 방법과 유사한 것으로 보고 되었으나[2], 상태 천이 확률의 재조정이 경험적인 방법에 의존한다는데 문제가 있다.

음성 신호의 시간축 변환 (Time Scale Modification; TSM) 은 음성을 구성하는 요소인 성도 전달 함수의 특성, 여기 신호의 특성, 피치 (pitch) 등의 특성을 유지한 채, 단지 발성 속도만을 변화시키는 기법으로, 주어진 음성을 마치 천천히 또는 빠르게 발성하는 것처럼 들리도록 변환하는 기술을 말한다. 이와 같은 시간축 변환의 응용 분야로서, 음성 메일링 시스템, 저전송률 부호화기 등을 들 수 있다[4]. 시간축 변환의 대표적인 알고리즘인 SOLA (Synchronized OverLap and Add) 방법[5]는 비교적 적은 계산만으로, 고음질의 시간축 변환된 음성을 얻을 수 있는 기법이다. 만일 고속의 음성이 입력되었을 때 SOLA 를 이용하여 천천히 발성하는 것처럼 들리도록 바꾸고, 이를 음성 인식기에 입력한다면, 마치 정상 속도로 발성된 음성이 인식되는 것과 같은 효과를 기대할 수 있을 것이다. 본 논문에서는 이와 같은 가정을 바탕으로 빠른 발화 속도의 음성에 대해 SOLA기법을 전처리 (preprocessing) 과정으로 사용하여, 빠른 발화 속도에 대해 강인성이 부여된 자동 음성 인식기를 제안하였다.

발화 속도를 반영한 자동 음성 인식기의 설계 시 고려되어야할 문제 중 하나는 주어진 음성으로부터 자동적으로 발화 속도를 추정해야 한다는 것이다. 이는 현재 입

력된 음성이 빠르게 발성한 음성이라고 판정된 경우에만 시간축 변환을 적용하는 것이 타당하기 때문이다. 주어진 음성으로부터 자동적으로 발화 속도를 추정하는 방법은 크게 HMM을 기반으로 하는 방법[1][6]과 음향적인 특징 변수들로만 추정하는 방법[7][8]으로 나눌 수 있다. HMM을 기반으로 하는 방법은 인식 결과와 함께 부가적으로 얻어지는 각 모델의 지속 시간, 모델내의 상태 지속 시간 등의 정보를 이용하는 방법으로, 단위 시간 내의 모델 수, 모델 지속 시간의 평균값 등을 계산하여, 음성의 발화 속도를 추정한다. 이와 같은 방법은 인식된 결과 자체가 오류를 포함할 수 있고, 모델의 경계 또한 오류를 포함할 수 있으므로, 추정된 발화 속도의 정확도가 사용된 음성 인식기의 성능에 의존적이다.

음향적인 특징 변수만을 이용하는 방법으로 신경회로망 (Neural network) 을 이용하여 사용된 음향 변수의 시상수 (time constant) 추정하는 방법[7], 혼합 가우시안 기법 (mixture gaussian) 을 이용한 방법[8] 등이 제안되었다. 신경회로망을 이용한 방법은 특징변수와 발화 속도와의 관계를 비선형적인 대응관계로 모델링 하는 방법으로, 신경망의 학습 시 사용되는 역전파 알고리즘 (back-propagation) 이 상대적으로 긴 수렴 시간을 요하는 경우, 계산량이 많아진다는 단점이 있다. 혼합 가우시안 방법은 주어진 음향적 특징 변수를 정상, 저속, 고속의 3가지 클래스로 분류 (classification) 하는데, 혼합 가우시안 함수를 사용하여 특징 변수가 어느 클래스에 속하는지를 확률적으로 나타내었다.

본 연구에서는 HMM을 기반으로 하는 음성 인식기를 사용하였으므로, 구현상의 용이함을 위해 HMM을 기반으로 한 방법을 적용하였다. 본 논문에서는 앞서 언급한 인식 오류가 발화 속도 추정에 어떠한 영향을 끼치는가를 알아보기 위해 약 1천개의 문장을 이용하여 정확한 음소 정보가 주어진 경우의 발화 속도와 인식된 음소 정보로부터 구한 발화 속도가 어떠한 차이를 나타내는지 살펴보았다. 또한, 이 결과를 바탕으로, 발화 속도의 추정 시 어떠한 변수를 사용하는 것이 타당한지를 경험적으로 제시하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 본 논문에서 사용한 발화 속도 추정 방법 및 유용성 여부, 발화 속도에 따른 음성 인식의 성능 변동을 제시하였다. 3장에서는 음성 신호의 시간축 변환 기법에 대해 간단히 설명하고, 제안된 음성 인식 시스템의 전체적인 구조를 소개하였다. 4장에서는 실험 결과를 통해 기존 기법과의 성능을 비교하였으며, 마지막으로 5장의 결론으로 본 논문을 끝맺었다.

II. 발화 속도의 추정과 음성 인식에의 영향

2.1. 발화 속도의 추정

음성 신호의 발화 속도에 따른 음성 인식의 성능 변화를 살펴보기 위해서는 먼저 발화 속도를 정량화하는 방법이 제시되어야 한다. 본 논문에서 적용한 HMM을 기반으로 하는 방법은 비터비 디코딩 (Viterbi decoding) 을 통해 얻어지는 각 모델의 시작시간 및 종료시간으로 모델의 길이를 구하고, 이들 모델 길이로부터 발화 속도를 추정하는 방법이다. 이 방법은 HMM을 기반으로 하는 음성 인식 장치에서 추가적인 계산 없이 간단하게 발화 속도를 추정할 수 있다는 장점이 있다.

주어진 모델 경계 정보로부터 발화속도를 추정하는 방법으로 다음의 두 가지 방법이 제안되었다. 첫 번째로 역 평균길이 (Inverse Mean Duration; IMD)로, 다음과 같이 정의한다[2].

$$IMD = \frac{n}{\sum_{i=1}^{n} d_i} \tag{1}$$

여기서 n은 발화 속도를 추정하고자 하는 문장 속에 포함된 전체 모델의 개수를 나타내며 d_i는 i번째 모델의 길이 (duration)을 나타낸다. 윗 식의 역수는 음소 길이 평균값임을 알 수 있으며, 따라서 빠르게 발성하는 음성에서는 모델의 길이가 짧으므로, IMD는 속도가 증가함에 따라 큰 값을 갖게 된다.

다음으로 사용되는 방법이 평균율 (Mean Rate; MR) 이며, 이는 다음과 같이 정의된다[2].

$$MR = \frac{\sum_{i=1}^{n} r_i}{n} \tag{2}$$

여기서 r_i는 i번째의 모델에 대한 길이의 역수 ($=\frac{1}{d_i}$)를 나타낸다. MR의 경우도 IMD와 마찬가지

로, 빠르게 발성하는 음성에 대해서는 각 모델의 r_i가 증가하므로, 발성 속도 증가에 비례적인 관계를 갖는다.

발화 속도를 추정하는 방법이 결정되면, 이제 경계 구분에 사용되는 모델을 어떻게 선택할 것인 지를 결정해야 한다. 여기서 말하는 "모델" 이란 음성 인식의 단위로, 문장, 단어, 음절, 음소, 음소를 구성하는 상태(state) 등이 될 수 있다. 예로서, 모델을 단어로 설정하는 경우 윗 식의 d_i는 문장 내 i번째 단어의 길이를 나타내며, 음소로 설정하는 경우에는 i번째 음소의 길이를 나타낸다. Morgan 등의 연구에서 모델 단위를 단어, 음소, 상태 등으로 변경시켜가며 수작업 레이블링 된 음성 신호에서 구한 IMD와 MR을 HMM을 이용한 자동 레이블링 결과로 구한 IMD, MR과 비교하였는데, 음소를 모델 단위로 사용한 경우에 두 값이 가장 유사하게 얻어진다고 보고하였다[2]. 이는 음소를 모델의 경계 구분에 사용하는 것이 추정된 발화 속도가 실측값과 가장 유사함을 의미하는 것으로, 본 논문에서도 음소를 발화 속도 추정의 모델 단위로 사용하였다.

다음으로 고려해야할 문제는, 음성 인식의 과정에서는 주어진 문장에 대한 음소열 (phoneme transcription)을 미리 알 수 없기 때문에, 음소 경계의 추정 시 HMM 정렬 (alignment)을 수행할 수 없다는 점이다. 이는 음소의 지속 시간이 모두 인식된 음소열을 바탕으로 추정되고, 100%의 인식율이 보장되는 인식기를 사용하지 않는 한, 음소열 자체에 오류가 발생되어 음소 경계에도 오류가 포함될 수 있음을 의미한다. 이러한 문제를 분석하기 위해, 본 논문에서는 다음과 같은 실험을 수행하였다.

먼저 학습 데이터로부터 각 음소에 대한 HMM을 생성한다. 학습 데이터에는 음성 신호의 파형뿐이 아니고, 각 음성에 대한 음소열을 모두 포함하고 있다. 학습과정에서 생성된 각 음소의 HMM을 이용하여 두 가지 방법을 이용하여 IMD, MR을 구한다. 첫 번째 방법은 주어진 음소열과 각 음소의 HMM을 이용하여, HMM 정렬 과정을 통해 음소의 경계를 추정, 이로부터 IMD, MR을 얻는 것이다. 두 번째 방법은 음소열이 주어져 있지 않은 상태에서 일반적인 음성 인식의 과정, 즉 비터비 디코딩을 통해 음소 경계를 얻고, 이로부터 IMD, MR을 계산하는 것이다. 즉, 첫 번째 방법은 오류가 포함되지 않은 올바른 음소열이 주어진 경우이며, 두 번째 방법은 인식된 음소열을 사용하는 경우이다.

발화 속도의 추정 시 IMD 와 MR의 유용성을 알아보기 위해 두 경우에 대한 상관 분포 (correlation distribution)을 구했으며, 그 결과가 그림 1과 2에 제시되었다. 그림에서 가로축은 HMM 정렬, 즉 올바른 음소열 에서 구한 IMD 또는 MR 값을, 세로축은 인식된 음소열 에서 구한 값을 나타낸다. 따라서 $x=y$축 상에 분포가 집중할수록 두 값 간의 유사성이 높음을 의미한다. 그림에서 보면, IMD 값이 MR값과 비교하여 $x=y$ 축 상에 집중하여 분포함을 알 수 있다. 이는 IMD값이 음소열에 오류가 포함되더라도, 올바른 음소열 에서 구한 값과 비교적 유사하게 얻어짐을 나타낸다.

본 논문에서는 빠르게 발성되는 음성에 대한 발화 속도의 추정 또한 중요하므로, 화자에게 특별히 빠른 속도

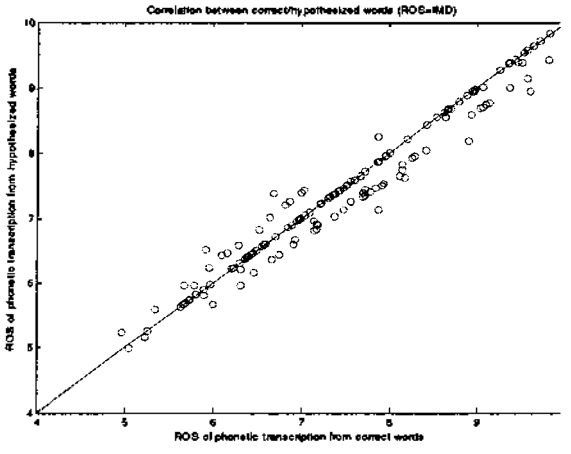

그림 1. 음소열이 주어진 경우와 인식된 음소열을 사용하는 경우의 IMD 상관 분포

Fig. 1. Correlation between IMDs computed from correct phoneme transcription and computed from recognized phoneme transcription

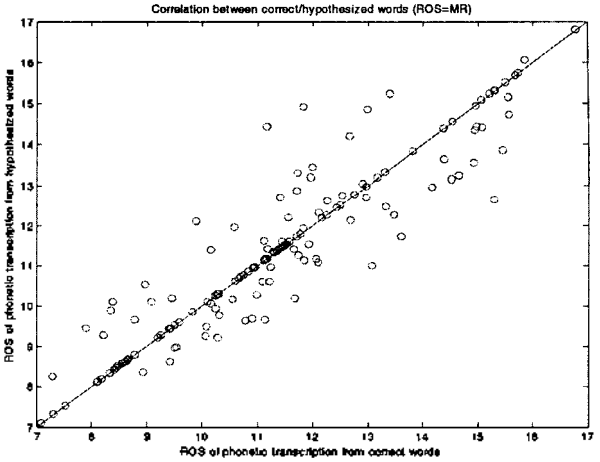

그림 2. 음소열이 주어진 경우와 인식된 음소열을 사용하는 경우의 MR 상관 분포

Fig. 2. Correlation between MRs computed from correct phoneme transcription and computed from recognized phoneme transcription

로 발성하도록 요구한 뒤, 취득된 음성에 대해서도 동일한 실험을 수행하였다. 그 결과가 그림 3과 4에 제시되었는데, 정상 속도의 음성에 대한 결과 보다는 다소 분산되어 분포함을 알 수 있다. 그러나, 정상 속도에 대한 실험 결과와 마찬가지로, IMD값의 경우가 분산 정도가 비교적 덜 함이 관찰된다. 이와 같은 실험 결과를 종합할 때, 올바른 음소열이 주어져 있지 않은 실제 환경에서는 IMD가 발화 속도의 추정에 더 유용한 정보를 제공한다고 볼 수 있다. 따라서 본 논문에서는 IMD를 발화 속도의 측정에 사용하였다.

2.2. 발화 속도에 따른 음성 인식의 성능

앞서 제시한 IMD를 이용하여 몇몇 음성에 대한 발화 속도와 인식율을 비교하였다. 여기서 사용된 음성은 휴대폰 전화기의 10자리 연속 숫자음 으로, HMM의 생성 조건 및 기타 실험 환경에 대해서는 4장에 제시하였다.

그림 5에 발화속도 (IMD로 표현됨) 대 인식율이 나타나 있다. 그림의 가로축은 학습 데이터에 포함된 전체 음성의 평균 IMD값과 테스트 데이터에 포함된 전체 음성의 IMD 비(ratio) 를 나타낸다. 즉 x값이 1보다 클수록, 테스트 음성이 학습 음성과 비교하여 빠르게 발성한 음성임을 나타낸다.

그림에 나타나있듯이, 음성 인식율은 발화 속도가 증가함에 따라 감소되는 것을 알 수 있으며, IMD의 비가 1인, 학습 데이터와 동일한 발화 속도의 음성에 대한 인식율 92.26%와 비교하면 최대 20% 저하된 인식율을 나타내고 있다. 이와 같은 저하된 인식율은 Morgan 등의 연구에서 밝힌바와 같이, 음소의 탈락, 특징 변수의 상이성 등에 그 원인이 있으며, 본 논문의 실험을 통해서도, 고속의 발화 음성에 대해 음소 탈락 오류(phoneme deletion error)가 유의하게 증가됨을 알 수 있었다.

III. 시간축 변환과 결합된 음성 인식 시스템

3.1. 음성 신호의 시간축 변환

고속의 발화 음성에 대한 인식율 저하를 방지하기 위해서는 주어진 음성을 본래의 발화 속도로 변환시키는 과정이 필요하다. 이때 발화 속도를 변화하더라도, 음성이 가지고 있는 본래 특성, 즉, 피치 (pitch), 성도전달함수 (vocal-tract response), 에너지 등의 정보는 그대로 유지되어야 한다. 이와 같은 조건을 만족하는 기법으로 음성 신호에 대한 시간축 변환 기법 (time scale modification)[4,5,9]이 사용될 수 있다. 시간축 변환은 음성의 여러 특징 변수들의 시간적 변화율을 변환시키는 기법으로, 변환된 음성은 본래의 음성과 비교하여 빠르게 혹은 느리게 발성하는 것처럼 들리게 된다.

초기의 시간축 변환은 단구간 퓨리어 변환 (short-time Fourier transform) 을 기반으로 구현되었으며, 대표적인 방법으로 Griffin 과 Lim에 의해 제안된 최소 자승 오차 변형 퓨리어 크기 추정[9](Least Square Error Estimate Modified Short Time Fourier Transform Magnitude; LSEE-MSTFTM) 을 들 수 있다. 이 방법은 시간축으로 변환될 신호와 원 신호의 퓨리어 변환 크기의 차이를 최소화 하도록 반복적으로 신호를 예측하여

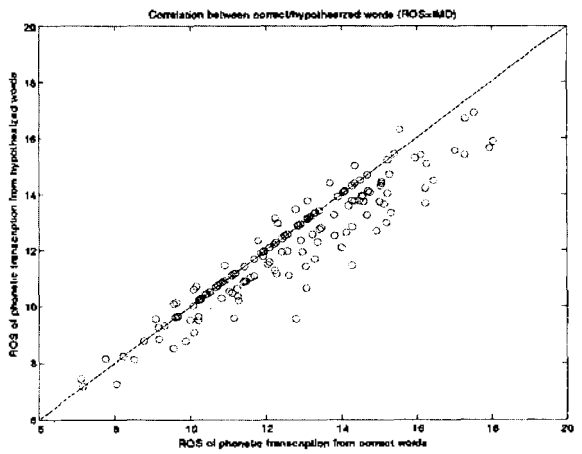

그림 3. 고속 발화 음성에 대한 IMD 상관 분포
Fig. 3. Correlation between IMDs computed from correct phoneme transcription and computed from recognized phoneme transcription for fast rate speech

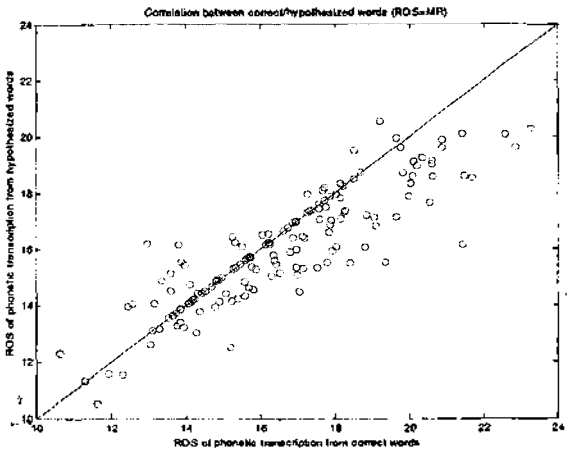

그림 4. 고속 발화 음성에 대한 MR 상관 분포
Fig. 4. Correlation between MRs computed from correct phoneme transcription and computed from recognized phoneme transcription for fast rate speech

시간축 변환된 음성을 얻는 방법이다. LSEE-MSTFTM이 반복 계산에 따른 계산량의 증가로 실시간 구현시 문제로 지적됨에 따라 보다 빠른 계산량 으로 고음질의 시간축 변환된 음성을 얻는 방법이 제안되었는데, Roucos에 의해 제안된 SOLA(Synchronized OverLap and Add) 방법[5]가 대표적이다.

SOLA는 기존의 방법이 분석 프레임 이동길이와 합성 프레임 이동길이가 다름으로써 발생하는 이질적인 펄스를 동기화(synchronized) 과정을 통해 제거하고 신호를 합성하는 방식이다. 즉, 그림 6에 제시한 바와 같이, 분석 프레임의 이동 길이가 S_a 이고 합성 프레임의 이동 길이가 S_s인 경우 이를 그대로 겹쳐 더한 경우 이질적인 펄스가 발생될 수 있으나, 그림에 나타난 바와 같이 k_m만큼 프레임을 이동시켜 펄스 위치를 맞추고 이질적인 펄스 없이 신호를 합성한다. 여기서 분석 프레임의 이동 길이를 합성 프레임의 이동 길이로 나눈 값을 시간축 변환의 정도 (factor)로 정의하는데, 이 값이 1보다 작은 경우는 빠르게 발성한 음성으로 변환하고, 1보다 큰 경우에는 느리게 발성한 음성으로 변환된다. 합성 프레임을 이동시키는 길이 k_m은 아래의 상호상관계수(correlation coefficient)가 최대로 되는 지점으로 선택된다. 즉,

$$k_m = \arg \max_k R(k) \tag{3}$$

여기서 $R(k)$는 이미 합성된 음성 $y(n)$과 분석 프레임내의 샘플값 $x(n)$의 상호 상관값으로 아래와 같이 주어진다.

$$R(k) = \frac{\sum_{j=0}^{L-1} y(mS_s + k + j)x(mS_a + j)}{[\sum_{j=0}^{L-1} y^2(mS_s + k + j) \sum_{j=0}^{L-1} x^2(mS_a + j)]^{1/2}} \tag{4}$$

여기서 L은 프레임의 길이를 나타낸다.

SOLA 방법은 비교적 간단한 방법으로 시간축 변환을 구현할 수 있으며, 변환음의 품질이 명료성과 자연성면에서 본래의 음성과 크게 떨어지지 않아 많은 용도에 사용되고 있다[8]. 이러한 사실은, SOLA 기법이 MFCC와 같은 음성 인식의 변수를 크게 변동시키지 않으면서 빠르게 발성하는 음성을 단지 느리게 들리도록 변환시키는데 이용될 수 있음을 의미하고, 이는 고속 발화음의 인식율 저하 문제를 해결하는데 이용될 수 있음을 나타낸다. 다음 절 에서는 지금까지 제시한 발화 속도 추정 기법, 시간축 변환 기법을 종합적으로 적용하여 고속 발화음의 인식율 저하를 극복하는 기법에 대해 소개하기로 한다.

3.2. 시간축 변환과 결합된 음성 인식 시스템

제안된 음성 인식 시스템의 구조를 그림 7에 제시하였다. 먼저 학습 과정에서는 정상 속도로 발성한 음성을 이용하여 음소별 HMM을 생성한다. 여기서 생성된 HMM을 이용하여 1차적으로 음성 인식을 수행한다. 인식된 결과를 토대로 IMD를 계산하여 발화 속도를 구하고, 이 값에 따라 시간축 변환을 선택적으로 수행한다. 즉, 현재 입력된 음성이 빠르게 발성된 음성으로 인식되면 시간축 변환을 적용하여 정상 속도의 음성으로 들리도록 변환 시키고 그렇지 않은 경우는 1차적으로 인식된 결과를 최종 인식 결과로서 출력하게 된다.

시간축 변환된 음성에 대해서는 2차적인 음성 인식을 수행하는데, 이 때 사용되는 HMM은 1차 인식때 사용한 것과 동일하다. 따라서 제안된 기법은 빠른 발성음에 대비한 별도의 HMM을 필요로 하지 않는다.

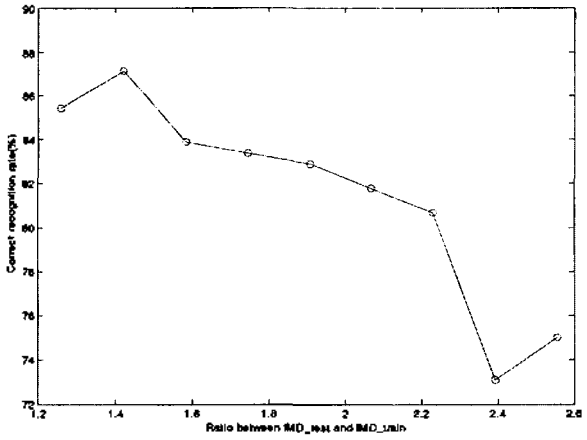

그림 5. 발화 속도(IMD)에 따른 인식율
Fig. 5. Recognition rate according to rate of speech (IMD)

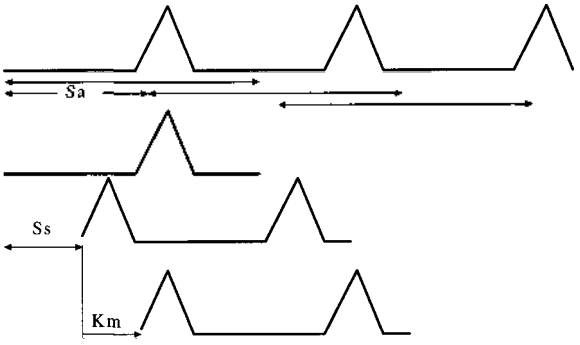

그림 6. SOLA 방법
Fig. 6. SOLA Method

표 1. HMM의 학습 조건
Table 1. HMM training coditions.

MFCC 차수	13 (에너지값 포함)
분석 프레임 이동 길이	10 msec
분석 프레임 길이	25 msec
HMM 종류	연속 모델
HMM 형태	좌-우 모델 (left-right model)
HMM 상태(state) 수	5 (짧은 묵음은 3)
혼합 가우시안의 개수	5

표 2. 음성 데이터에 따른 인식율과 평균 발화 속도
Table 2. Recognition rate and average rate of speech according to speech data.

음성 데이터	인식율 (%)	평균발화속도 (IMD)
학습 데이터	96.59	8.21
정상속도의 테스트 데이터	92.26	10.37
빠른속도의 테스트 데이터	83.32	14.42

이와 같은 시스템을 구현하기 위해서는 최적의 인식율을 가져오는 시간축 변환 정도와 시간축 변환을 선택하는 기준이 결정되어야 한다. 본 논문에서는 반복적인 실험과 결과를 통한 경험적인 방법으로 결정되도록 하였으며, 이에 대한 자세한 결과는 4장에서 제시하도록 한다.

IV. 실험 및 결과

본 논문에서 제안된 기법의 성능 평가를 위해 실제 고속의 발화음에 대해 음성 인식을 수행하고 결과를 살펴보았다. 음성 인식의 대상은 이동 통신 전화번호인 10자리 숫자음을 사용하였으며, 각 숫자음은 트라이폰 부모델(triphone subword model)이 연결된 형태로 나타내었다. 사용된 부모델의 개수는 총 24개인데, 이중 2개의 부모델은 통계적인 방법에 따라 공유 모델(shared-model)이 사용되어 실제적으로는 22개의 HMM이 사용되었다. 음성 파라메터는 13차 MFCC(Mel-Frequency Cepstral Coefficients) 가 이용되었으며, 델타 값 및 델타-델타 값을 포함하는 총 39개의 변수가 HMM의 생성과 음성 인식에 사용되었다. 이들 변수의 계산 및 학습 조건은 표 1에 제시하였다.

10자리 전화번호는 16명의 화자에 의해 10종류의 전화번호를 10번 반복 발성하여 녹음 하였는데, 동일 화자 내에서의 음성 변동을 고려하기 위해 충분한 시간적 간격을 두고 반복 녹음하였다. 또한 빠르게 발성하는 음성

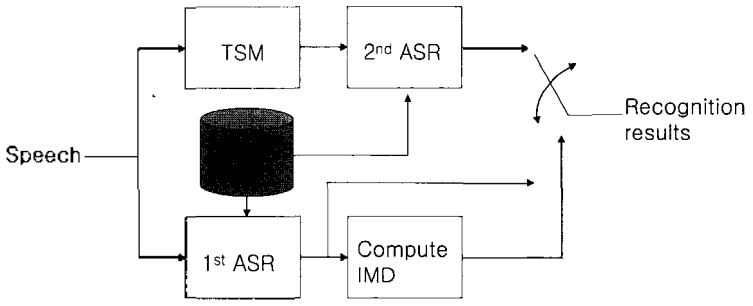

그림 7. 시간축 변환 기법이 포함된 음성 인식 시스템의 블럭도
Fig. 7. Block diagram of an automatic speech recognition system with time scale modification technique

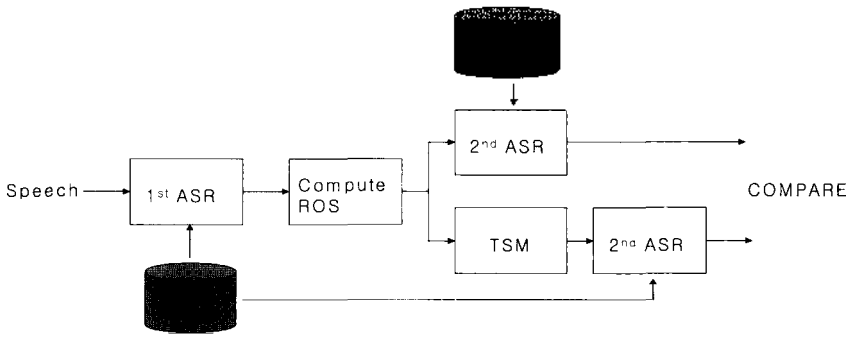

그림 8. 시간축 변환 음성의 인식율에 대한 실험 블럭도
Fig. 8. Experiment design for recognition rate of the time-scaled speech signals.

에 대한 성능을 평가하기 위해 동일한 화자들로부터 동일한 전화번호를 10회 빠르게 발성한 음성을 추가적으로 취득하였다. 여기서 얻어진 음성에 대해서는 실제 빠르기를 알아보기 위해 2장에서 제시한 IMD를 취득한 문장마다 계산하였으며, 정상 속도로 발성한 음성과 비교하여 30%이상의 속도차이를 나타내지 않는 경우에는 화자에게 좀더 빠른 속도로 발성할 것을 요구하여 재녹음하였다.

HMM의 학습에는 8명의 화자가 정상 속도로 발성한 10종류 전화 번호 10개의 반복 음성이 사용되었는데, 이는 총 800개의 문장, 9600개의 단어에 해당한다. 테스트 데이터의 구성에는 학습시 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.

또한 시간축 변환이 결합된 음성 인식의 기법과 기존 방법간의 성능 비교를 위해, 본 논문에서는 고속의 발화 음성으로 별도의 HMM을 생성하여 음성 인식에 사용하는 방법[2]을 비교 대상으로 삼았으며, 이를 위해 학습 데이터에 포함된 8명의 화자들이 고속으로 발성한 800개의 문장들을 이용하여 별개의 HMM을 생성하였다.

4.1. 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 (IMD)와 단어 인식율이 나타나있다. 발화 속도와 인식율 간의 관계를 정확히 알아보기 위하여 표 2 에 제시한 IMD 값은 음소열 정보가 미리 주어진 상태에서, HMM 정렬을 통해 얻은 음소의 경계 정보로 구하였다.

학습 데이터와 정상 속도의 테스트 데이터 간 인식율은 각각 96.6% 와 92.3% 두 데이터간의 유의한 차이는 관찰되지 않았다. 두 데이터군의 평균 발화 속도는 각각 8.21 과 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 83.3%로 정상 속도의 발성음에 비해 9% 낮은 인식율을 나타내었다. 발화 속도면 에서는 10.37 대 14.42 로 빠른 발성음이 25% 정도 높은 IMD 값을 나타내었다.

4.2. 음성 신호의 시간축 변환이 전처리로 사용된 경우의 인식율

빠르게 발성된 음성에 대한 시간축 변환 적용의 타당성을 알아보기 위해, 고속 발성음에 대해 시간축 변환이

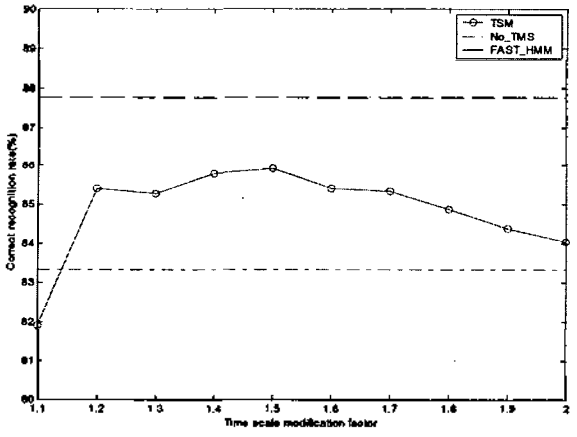

그림 9. 시간축 변환 정도에 따른 음성 인식율
Fig. 9. Recognition rate according to time scale modification factor.

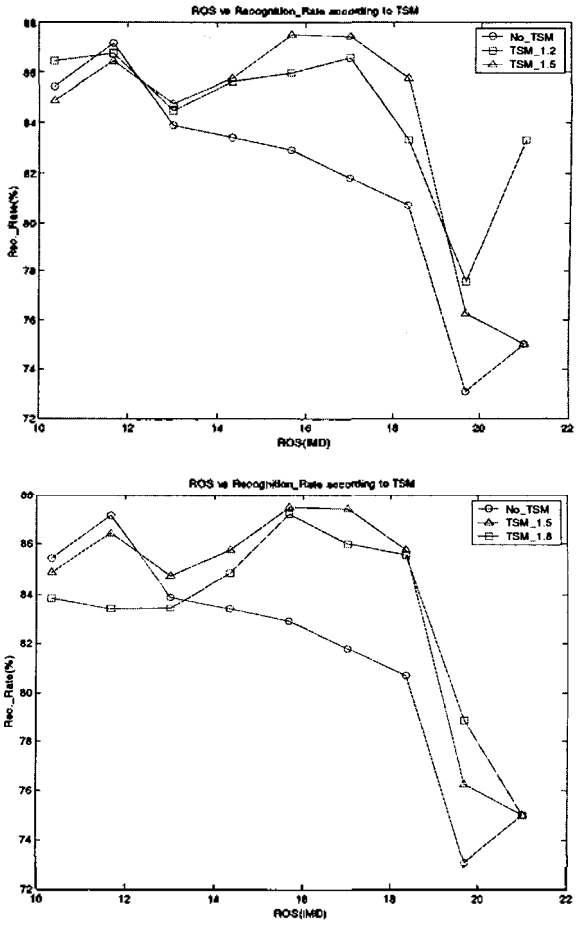

그림 10. 시간축 변환 정도에 따른 발화 속도 와 인식율 관계
Fig. 10. Relationship between rate of speech and recognition rate according to time scale modification factor.

적용한 후, 인식율을 구하고 시간축 변환이 사용되지 않은 경우와 비교하였다. 또한 고속 음성 데이터들로 생성된 HMM을 이용하여 인식하는 경우와 성능을 비교하였다. 전체적인 실험 과정이 그림 8에 제시되었다.

실험의 첫 번째 시간축 변환의 정도에 대한 인식율이 조사하였으며, 그 결과가 그림 9에 제시되었다. 실험은 고속의 발화음을 대상으로 하였으므로, 시간축 변환의 정도도 항상 1보다 높은 경우로 한정하였다. 비교의 용이성을 위하여, 그림 9에는 시간축 변환을 적용하지 않은 경우의 인식율 (세로축의 83.32 값과 만나는 직선)과 고속 음성으로 생성된 HMM을 사용하여 얻은 인식률 (최상단의 직선) 를 함께 표시하였다.

시간축 변환의 정도에 따른 인식율의 변화를 보면, 1.1배의 시간축 변환에 있어서는 시간축 변환을 수행하지 않은 음성보다 오히려 낮은 인식율을 나타내었으며, 1.1배 이상의 변환 정도에서는 상승되는 인식율을 보이다가 1.5배를 기준으로 완만하게 하강함이 관찰되었다. 일반적으로 시간축 변환이 적용된 음성은 변환 정도가 작은 경우에는 음질적인 저하나 부자연성이 거의 느껴지지 않지만, 변환의 정도가 커질수록 부자연성이 증가된다고 알려져 있다[8]. 이러한 부자연성의 증가는 음성 인식의 변수인 MFCC에도 영향을 줄 것으로 예상되며, 따라서 1.5배 이상의 시간축 변환에서 인식율의 성능이 저하되는 주된 이유는, 과도한 시간축 변환에 따른 변환음의 품질 저하 및 이에 따른 MFCC 값의 왜곡으로 볼 수 있다.

작은 시간축 변환 정도 (1.1배)와 큰 시간축 변환 정도에서의 인식율 저하는 변환음의 발화 속도 (IMD)와 학습 데이터의 발화 속도 차이에 원인이 있는 것으로 보인다. 이론적으로는, 시간축 변환된 음성의 발화 속도가 학습

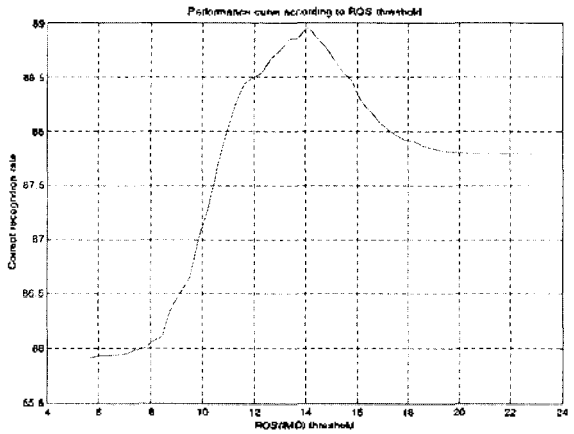

그림 11. 발화 속도의 임계치에 따른 인식율
Fig. 11. Recognition rate according to threshold of rate of speech.

데이터의 발화 속도와 정확히 일치 되었을 때 최고의 인식율이 얻어져야 하나 실제 실험 결과를 보면 최고의 인식율이 얻어지는 1.5배 시간축 변환에서는 9.85의 IMD가 얻어졌고, 학습 데이터의 IMD는 전술한 바와 같이 IMD=8.21 로서, IMD가 정확히 일치하지는 않았다. 학습 데이터의 발화 속도와 가장 유사하게 되는 경우는 시간축 변환을 1.8배로 적용하는 경우로서, IMD=8.31가 얻어졌다. 1.8배 적용시 최고의 인식율을 얻지 못한 이유는 앞서 언급한 바와 같이, 과도한 시간축 변환에 따른 변환음의 품질 저하가 인식율의 저하 요인으로 작용되었기 때문이라 생각된다.

고속의 발성음에서 생성된 HMM을 사용하는 경우의 오류 감소율은 26.56% 로서, 시간축 변환 적용시 최대로 얻을 수 있는 오류 감소율 15.53%와 비교하여 좀 더 나은 성능을 나타내었다. 이때의 오류 감소율 차이는 고속 발성음과 정상 속도의 발성음의 차이중 시간축 변환에 의해 보상될 수 없는 성분을 반영한다고 해석할 수 있다. 반대로 시간축 변환에 의해 얻을 수 있는 15.53%의 오류 감소율은 시간축 변환으로 보상될 수 있는 성분의 크기로 간주하면, 시간축 변환에 의해 보상되는 성분이 그렇지 않은 성분보다 상대적으로 더 크다고 말할 수 있다.

본 논문에서 적용된 시간축 변환은 주어진 발성음의 시간축 길이를 확장하여, 결과적으로 향상된 시간 해상도를 갖는 MFCC를 제공하여 음성 인식의 성능을 향상시키는 방법으로 해석할 수 도 있다. 그러나 이와 같은 원리에 입각한 방법이라면, 굳이 시간축 변환을 사용하지 않더라도, MFCC 계산시의 프레임 레이트를 감소시켜 시간 해상도를 증가시키는 방법이 고려될 수 있다. 예로서 고속 발화음에 1.5배의 시간축 변환을 적용하고 10ms의 프레임 레이트로 MFCC를 계산하는 방법과, 시간축 변환을 수행하지 않고 7.5ms의 프레임 레이트로 MFCC를 계산하는 방법은 동일한 개수의 MFCC를 생성하게 된다.

본 논문에서는 이와 같이 단순히 MFCC의 시간 해상도를 증가시키는 방법과의 비교를 위해, 1.5배의 시간축 변환을 수행하고 10ms 프레임 레이트로 MFCC를 추출하는 방법과 시간축 변환 없이 프레임 레이트를 1.5배 증가시킨 방법간의 성능을 비교하였다. 두 방법 모두 동일한 학습 데이터에서 생성된 HMM을 사용하였다(프레임 레이트는 10ms). 오류 감소율에서 시간축 변환을 사용한 방법은 전술한 바와 같이 15.53% 를, 1.5배 증가된

프레임 레이트를 사용한 경우에 대해서는 14.15% 을 나타내어 두 방법간의 유의한 차이는 발견되지 않았다. 이는 시간축 변환이 MFCC의 시간축 해상도를 향상시키는 것과 음성 인식의 성능면에서 동일한 효과를 거둘 수 있음을 의미한다. 또한 고속 SOLA 알고리즘[10]을 사용하는 경우, MFCC의 추출 과정보다 비교적 적은 계산량이 소요되므로 제안된 방법은 계산량을 크게 증가시키지 않으면서 향상된 시간 해상도의 특징 변수를 제공하는 방법이라 볼 수 있다.

4.3. 발화 속도에 따른 적응적 시간축 변환

본 논문의 최종 목표는 주어진 음성의 발화 속도를 추정하고, 이에 따라 적절히 발화 속도를 보상하여 음성 인식을 수행하는 것이다. 이와 같은 목표를 달성하기 위해서는 발화 속도에 따라 적응적인 시간축 변환을 수행해야 한다. 적응적인 시간축 변환을 수행하는 한 가지 방법은 입력된 음성으로부터 발화 속도를 추정하고, 여기에 따라 시간축 변환의 정도를 달리하여 속도를 보상하는 것이다. 이를 위해 본 논문에서는 시간축 변환의 정도를 달리하면서, 발화 속도 (IMD)에 따른 인식율을 조사하였다. 그 결과가 그림 10에 제시되었는데, 발화 속도와 시간축 변환의 정도 간에 유의한 상관관계를 관찰할 수 없었으며, 1.5배의 시간축 변환에서 발화 속도와 무관하게 최대의 인식율이 얻어짐을 알 수 있었다. 따라서, 시간축 변환의 정도는 IMD 값에 관계 없이 1.5배 고정된 값을 사용하였다.

다음으로 고려할 사항은, 어떤 경우에 시간축 변환을 적용할 것인가 하는 것이다. 음성 인식 시스템에 입력되는 음성은 빠른 음성과 느린 음성이 모두 포함되므로, 만일 일률적으로 1.5배의 시간축 변환이 포함된다면 느리게 또는 정상적인 속도의 발성음에 대해서는 오히려 저하된 인식율을 나타낼 것이다. 이와 같은 사항을 고려하여 본 논문에서는 발화속도의 임계치에 대한 인식율을 조사하였다. 여기서 발화속도의 임계치라 함은 주어진 음성의 IMD값을 비교하는 척도로서, 임계치보다 높은 값이면 시간축 변환을 수행하고 그렇지 않으면 1차적으로 얻어진 음성 인식 결과를 최종 인식 결과로 그대로 출력하게 된다.

임계치에 대한 인식율을 그림 11에 제시하였다. IMD=14 근방에서 최대의 인식율이 얻어짐을 알 수 있다. 여기서 제시된 인식율이 4.3절에 제시한 인식율 보다 다소 증가되어 나타나는 것은, 고속의 발성음 뿐이 아니고 정상적으로 발성된 음성, 즉 시간축 변환을 수행하지 않은 음성에 대한 결과도 함께 포함되어 있기 때문이다. 결론적으로, 제안된 시스템은 IMD=14 이상의 음성에 대해 1.5배의 시간축 변환을 수행하여 음성 인식을 수행하였다.

V. 결 론

본 논문에서는 음성 신호의 변환 기법이 단순히 오락과 흥미를 유발하는 용도로 사용되는 것이 아닌, 음성 인식기의 성능을 향상시킬 수 있는 방법으로 적용될 수 있는 가능성을 제시하였다. 시간축 변환 기법이 적용되었는데, 주된 이유는 현존하는 다른 음성 변환 기법에 비해 고음질의 음성으로 변환이 가능하며, 비교적 적은 계산량이 요구되기 때문이다. 만일 현재의 피치 변경 기법 등이 보다 고품질의 변환음을 생성할 수 있다면, 이들 기법들을 다양하게 적용하여, 보다 향상된 인식율을 얻을 수 있을 것으로 기대된다.

제안된 기법은 HMM 기반의 음성 인식 기법에서 부가적으로 얻어지는 음소 경계 정보를 발화 속도 추정에 이용하였다. 따라서, HMM을 기반으로 하지 않는 음성 인식 기법이 사용되는 경우에는 음향적인 특징 변수로만 발화 속도를 추정하는 방법이 사용되어야 할 것이다.

실험을 통한 제안 기법의 성능 평가에서, 특별히 빠르게 발성한 음성에 대해서 15.53%의 에러 감소율을 나타내었다. 결론적으로, 제안된 방법은 대화체 문장 등에 자주 발견되는 고속의 발화 음성에 대해서도 적용 가능한 음성 인식 기법으로 사료된다.

감사의 글

본 연구는 학술진흥재단 2003년 신진교수과제 (D00321)에 의한 결과임.

참 고 문 헌

1. N. Mirghafori, E. Fosler, and N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes," The proceedings of EUROSPEECH95, pp. 491-494,

Madrid, Spain, September 1995.
2. N. Mirghafori, E. Fosler, and N. Morgan, "Towards robustness to fast speech in ASR," The proceedings of ICASSP96, pp. 335-338, Atlanta, USA, 1996.
3. J. Zheng, H. Franco, and A. Stolcke, "Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition," Speech Communication, **41**, pp. 273-285, 2003.
4. J. Makhoul and A. E. Jaroudi, "Time-scale modification in medium to low rate speech coding," proc. of ICASSP, 1, pp. 1705-1708, 1986.
5. S. Roucos and A. M. Wilgus, "High quality time-scale modification for speech," proc. of ICASSP, 1, pp. 493-469, 1985.
6. M. J. Russell, K. M. Ponting, and M. J. Tomlinson, "Measure of local speaking-rate for automatic speech recognition," IEE Electronics Letters, 35(10), pp. 787-789, 1999.
7. M. H. Nguyen and G. W. Cottrell, "A technique for adapting to speech rate," The proceedings of the 1993 IEEE-SP workshop, 6-9, pp. 382-391, September 1993.
8. R. Fallthauser, T. Pfau and G. Ruske, "On-line speaking rate estimation using Gaussian mixture models," The proceedings of ICASSP2000, pp. 1355-1358, 2000.
9. D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," IEEE Trans. on Acoust., Speech, Signal Processing, **ASSP-32**, pp. 236-243, 1984.
10. S. Yim and B. I. Pawate, "Computationally efficient algorithm for time scale modification(GLS-TSM)," proc. of ICASSP, 2, pp. 1009-1012, 1996.

저자 이력

• **이 기 승 (Ki-Seung Lee)**

1968년 1월 25일 생.
1991년 2월: 연세대학교 전자공학과(공학사)
1993년 2월: 연세대학교 대학원 전자공학과(공학석사)
1997년 2월: 연세대학교 대학원 전자공학과(공학박사)
1997년 3월~1997년 9월: 연세대학교 신호처리 연구센터 선임 연구원
1997년 10월~1999년 8월: AT&T Shannon Lab, Consultant
1999년 9월~2000년 9월: AT&T Shannon Lab, Senior Technical Staff Member
2000년 11월~2001년 8월: 삼성종합기술원 HCI Lab 전문연구원
2001년 9월~현재: 건국대학교 정보통신 대학 전자 공학부 조교수
※주관심 분야: 음성 합성, 운율 제어, 음성 변환, 음성 부호화기 등.