2001年度 第14回 信號處理合同學術大會 論文集 第14卷 1號

휴대폰 단말기에 적용을 위한 강인한 음성인식

손 종 목, 정 성 윤, 배 건 성
경북대학교 전자·전기공학부

Robust Speech Recognition for Application to Mobile Phone

Jong Mok Son, Sung Yun Jung, Keun Sung Bae
School of the Electronic & Electrical Engineering, Kyungpook National University
sjm@palgong.knu.ac.kr

요 약

최근 음성인식이 인간과 기계 사이의 자연스러운 통신을 위한 가장 중요한 수단으로 인식되어 이와 관련된 연구가 꾸준히 이루어져 왔으며, 일부 응용 분야에서는 성공적으로 적용되고 있다. 하지만, 좀 더 다양한 응용분야에 적용하기 위해서는 실제 환경에 존재하는 여러 가지 주변잡음에 강인한 특성을 가지는 인식 시스템이 요구된다.

본 연구에서는 음성인식 시스템을 휴대전화에 적용하기 위해 도메인 적응 기법, LDA (Linear Discriminant Analysis) 기법 등을 도입하여 시스템 DB의 크기를 줄이고 잡음에 대한 강인성을 높이고자 하였으며, HMM (Hidden Markov Model)에 기반한 음성인식 시스템을 사용하여 각 기법의 적용에 따른 인식성능을 평가하였다.

Ⅰ. 서 론

최근 음성인식이 인간과 기계 사이의 자연스러운 통신을 위한 가장 중요한 수단으로 인식되어 이와 관련된 연구가 꾸준히 이루어져 왔으며, 일부 응용 분야에서는 성공적으로 적용되어 다양한 제품들이 출시되고 있다. 하지만, 좀 더 다양한 응용분야에 음성인식 기술을 적용하기 위해서는 실제 환경에 존재하는 여러가지 주변 잡음에 강인한 특성을 가지는 인식 시스템이 요구된다. 일반적으로 음성인식시스템은 훈련과 인식시의 주변환경이 다를 경우 인식성능이 크게 저하되는 것으로 알려져 있다. 만약 인식기가 사용될 때의 주변환경을 미리 알고 있을 경우, 주변환경을 고려한 훈련을 통하여 음성신호의 왜곡에 의한 인식성능 저하를 상당히 줄일 수 있다. 그러나, 휴대전화 환경에서는 주변환경이 다양하게 변화하여, 훈련 시 모든 상황을 충분히 고려할 수 없기 때문에 인식환경과의 차이를 어느 정도 보상해야 할 필요가 있다.

음성인식 시스템의 인식성능을 저하시키는 요인으로는 크게 부가잡음에 의한 신호의 왜곡과 채널 특성의 차이로 볼 수 있다. 부가잡음은 음성신호에 합의 형태로 나타나고, 채널간의 차이는 음성신호에 대해 컨벌루션의 형태로 나타난다. 때문에, 부가잡음에 의해 왜곡된 음성신호의 개선은 주파수 영역이나 웨이브렛 영역에서, 채널 특성의 차이는 CMN(Cepstral Mean Normalization) 등과 같이 켑스트럼 영역에서 연구가 많이 이루어졌으며[1,2], 훈련과 인식시의 환경 차를 보상하기 위한 적응기법도 많이 연구되어 왔다[3].

본 연구에서는 휴대전화가 사용되는 다양한 환경을 고려하여 도메인 적응기법, LDA (Linear Discriminant Analysis) 기반 기법 등을 적용하여 HMM에 기반한 음성인식 시스템의 성능저하를 줄이고자 하였으며, 실험을 통하여 각 기법들의 적용에 따른 인식성능의 변화를 알아보았다.

본 논문의 구성은 다음과 같다. Ⅰ장 서론에 이어 Ⅱ, Ⅲ장에서는 MAP(Maximum A Posteriori) 적응기법과 LDA 기법에 대해 각각 설명하고, Ⅳ장에서는 실험환경을 기술하며 휴대전화 환경에서 도메인 적응기법과

LDA 기반 기법을 적용하였을 경우의 인식결과를 제시하고, 결과를 검토한다. 마지막으로 V장에서 결론을 맺고 향후 연구방향을 제시한다.

II. Maximum A Posteriori Adaptation

새로운 잡음환경에 HMM 모델을 적응시킴으로써 잡음에 의한 인식성능 저하를 상당히 줄여줄 수 있음이 일반적으로 알려져 있다. 이를 위한 방법으로 MAP나 MLLR(Maximum Likelihood Linear Regression)과 같은 적응 방식이 사용될 수 있다. 충분한 적응 데이터가 주어질 경우 일반적으로 MAP 적응 기법이 MLLR보다 적응 성능이 좋은 반면, 작은 적응 데이터에 대해서 MLLR이 나은 성능을 나타내는 것으로 알려져 있다[4].

연구실 환경에서 훈련된 시스템에 대해서 휴대폰 환경으로의 모델적응을 위해 MAP를 사용하였다. 실험에서는 식 2-1을 사용하여 HMM 모델의 평균 값만을 적응하여 사용하였다.

$$\widehat{m_k} = \frac{\tau_k \mu_k + \sum_{t=1}^{T} c_{kt} x_t}{\tau_k + \sum_{t=1}^{T} c_{kt}} \qquad \text{(식 2-1)}$$

III. Linear Discriminant Analysis

N 개의 독립적인 벡터 $\{x_i\}_{1 \le i \le N}, x_i \in R^n$를 고려할 때, 각각의 벡터는 매핑 $l : \{1, \ldots, N\} \rightarrow \{1, \ldots, J\}$에 의하여 단지 1개의 클래스 $j \in \{1, \ldots, J\}$에 속하게 된다고 가정한다. 클래스 j가 평균 μ_j와 분산 Σ_j, 표본수 N_j에 의해 특징 지워진다고 할 때 식 3-1과 식 3-2의 가정은 유효하다.

$$\mu_j = \frac{1}{N_j} \sum_{i \in l^{-1}(j)} x_i \qquad \text{(식 3-1)}$$

$$\Sigma_j = \frac{1}{N_j} \sum_{i \in l^{-1}(j)} x_i x_i^T - \mu_j \mu_j^T \qquad \text{(식 3-2)}$$

단, $\sum_{j=1}^{J} N_j = N$ 이다. 클래스의 정보는 식 3-3, 식 3-4와 같은 2개의 scatter 행렬로 표시된다.

· Within-class scatter:

$$W = \frac{1}{N} \sum_{j=1}^{J} N_j \Sigma_j \qquad \text{(식 3-3)}$$

· Between-class scatter:

$$B = \frac{1}{N} \sum_{j=1}^{J} N_j \mu_j \mu_j^T - \overline{\mu} \overline{\mu}^T \qquad \text{(식 3-4)}$$

LDA의 목적이 식 3-5의 목적함수를 최대로 하는 선형 변환을 찾는 것이므로, 이는 식 3-6과 같은 일반화된 고유치 문제의 고유벡터를 구하는 것이 된다.

$$J(\theta) = \frac{|\theta B \theta^T|}{|\theta W \theta^T|} \qquad \text{(식 3-5)}$$

$$Bx = \lambda W x \qquad \text{(식 3-6)}$$

IV. 실험 및 검토

실험환경은 다음과 같다. 도메인 적응기법과 LDA를 인식시스템에 적용하였을 경우, 각 방법에 따른 인식률을 보기 위해 PTM(Phonetic Tied-Mixture) 모델을 사용하여 가변어휘 인식기를 구현하였다[5]. 인식시스템의 기본단위로는 한국어 자소에 기반한 50개의 유사음소를 사용하였다. 발음사전의 생성을 위해서 한국어 읽기 규칙에 기반하여 단어를 유사음소 열로 표기한 후 유사음소 사이에 나타나는 부가음운을 확률적으로 첨가하였다. PTM를 위해 10개의 클래스를 정의하였고, 하나의 음소를 모델링하기 위해 3 상태 Bakis(Left-to-Right) 모델을 사용하였으며, 기본 유사음소 중 그 길이가 짧은 유사음소를 나타내기 위하여 상태의 생략을 허용하였다. 각 상태에서 관측분포를 나타내기 위해 사용한 가우시안의 갯수는 8개이다. 이때, 기본 음향모델의 크기는 410 KBytes이며, LDA를 적용하였을 경우 269 KBytes이다.

기본 시스템의 훈련을 위해서 ETRI(Electronics and Telecommunications Research Institute)의 445 DB 중 훈련용 화자(남자 16명, 여자 14명)의 데이터를 8 KHz로 다운 샘플링하여 사용하였다. 각 기법의 적용에 따른 인식률의 변화를 보기 위한 평가용 DB로는 삼성 name DB를 사용하였으며, 기본 모델을 휴대폰 환경으로 적응시키기 위해 삼성 command DB를 사용하였다.

실험과정은 다음과 같다. 음성 데이터를 전처리 계수 0.95로 전처리 후, 20ms 길이의 해밍 윈도우를 10ms 마다 취하여 구간 분석하였다. 각 구간에서 1차의 에너지와 12차의 멜 캡스트럼을 구하고, 현재 구간을 포함

표. 4-1 음성 데이터 분석

Sampling Frequency	8 KHz
Quantization	16 bits
Hamming Window	20ms (160 points)
Frame Rate	10ms (80 points)
Feature Parameters	1 energy component 1 Δ-energy component 1 Δ^2-energy component 12 MFCC components 12 Δ-MFCC components 12 Δ^2-MFCC components

한 전후 각 6 구간(전체 13 구간)의 정보를 이용하여 1차의 차분 에너지와 12차의 차분 멜 캡스트럼, 그리고 차분-차분 값을 구하였다. 표 4-1에 음성의 분석 조건을 나타내었다.

휴대폰 환경으로의 적응을 위하여 MAP를 적용할 때, 전 모델에 대하여 τ_k를 10으로 고정시켜 사용하였다. LDA 기법의 적용에 있어서는 유사음소 모델의 각 상태를 구별하고자 하는 클래스로 정의하고 선형변환 행렬을 구하였으며, 구해진 특징 파라미터(39차)에 변환 행렬을 적용하여 프레임 당 24차로 관측벡터의 차수를 낮추었다.

표 4-2에 각 기법을 적용하였을 경우의 인식률을 나타내었다. 표 4-2는 적용 기법 외에 부가적인 기법을 인식시스템에 적용하지 않았을 경우의 결과이다. 결과를 살펴보면 훈련과 인식 시의 환경에 큰 차이가 있는 기본 시스템이 전체 82 %의 인식률을 보인 것에 비하여 MAP 기법을 적용하여 환경 차를 보상해 준 경우 전체 86.9 %로 인식률이 크게 향상됨을 볼 수 있다. 또한, MAP+LDA의 경우 87.99 %로 큰 향상을 볼 수 있다. 인식실험에 사용된 데이터에 따른 결과를 보면, 비교적 잡음이 심하지 않은 HHP SCH-600과 HHP SPH-7000을 사용하여 수집한 데이터의 경우 평균적으로 큰 차이가 없는 반면, 주변 잡음이 심하게 나타나는 HF SCH-600로 수집한 데이터의 경우 인식률이 크게 향상됨을 볼 수 있다.

V. 결 론

본 연구에서는 휴대전화가 사용되는 다양한 환경에서 강인한 음성인식을 위해 MAP 적응 기법과 LDA 기법을 도입하고 HMM에 기반한 음성인식 시스템의 인식

표 4-2. 각 기법에 따른 인식률비교(%)

	Base	MAP	MAP+LDA
HHP SCH-600	86.55	88.96	87.14
HHP SPH-7000	93.24	89.85	92.90
HF SCH-600	71.07	83.55	83.27
Total	82	86.90	87.99

성능을 평가하였다.

각 기법별로 인식성능을 평가해 본 결과 환경의 차이를 보상해 주었을 경우 인식성능에 큰 향상이 있었으며, LDA 기법을 적용하여 시스템 모델의 크기를 줄였음에도(410KBytes -> 269KBytes) MAP만을 적용하였을 경우보다 약간의 인식률 향상을 볼 수 있었다. 향후, 다양한 휴대폰 모델에 대한 실험과 함께 음성신호의 개선에 관한 연구가 휴대폰 환경에서 음성인식 기술을 적용하기 위해서 이루어져야 한다.

참고 문헌

[1] Alejandro Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," *Ph.D. thesis, Carnegie Mellon University*,1990.

[2] Fu-Hua Liu, Richard M. Stern, Xuedong Huang, Alejandro Acero, "Efficient Cepstral Normalization for Robust Speech Recognition," *Proc. of the sixth ARPA Workshop on Human Language Technology*, 1993.

[3] Rolf Bippus, Alexander Fischer, Volker Stahl, "Domain Adaptation for robust Automatic Seech Recognition in CAR Environments," *Eurospeech 99,* pp. 1943-1946, 1999.

[4] Xuedong Huang, Alex Acero, Hsiao-wuen Hon, *Spoken Language Processing*, Prentice Hall.

[5] Akinobu Lee, Tatsuya Kawahara, Kazuya Takeda and Kiyohiro Shikano, "A New Phonetic Tied-Mixture Model for Efficient Decoding," *International Conference on Acoustic, Speech and Signal Processing*, vol. 3, pp. 1269-1272, 2000.

[6] Michael L. Shire, "Data-Driven Modulation Filter Design under Adverse Acoustic Condition and Using Phonetic and Syllabic Units," *Eurospeech 99,* pp. 1123-1126, 1999.

[7] R. Haeb-Um bach and H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," *International Conference on Acoustic, Speech and Signal Processing*, vol. 1, pp.13-16, 1992.

[8] Markus Lieb and Reinhold Haeb-Umbach, "LDA derived Cepstral Trajectory Filters in Adverse Environmental Conditions," *International Conference on Acoustic, Speech and Signal Processing*, 2000.

[9] Jean-Luc Gauvain and Chin-Hui Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio processing*, vol. 2, no, 2, 1994.