Modified Bessel 함수 근사화를 적용한 MMSE STSA 전처리 기법의 음성인식 성능 비교

손 종 목, 김 민 성, 배 건 성
경북대학교 전자·전기공학부

Comparison of Recognition Performance for Preprocessing Method of MMSE STSA with Approximated Modified Bessel Function

Jong Mok Son, Min Sung Kim, and Keun Sung Bae
School of Electronics and Electrical Engineering, Kyungpook National University
sjm@palgong.knu.ac.kr

요 약

본 연구에서는 음성신호의 왜곡에 대해 음성 부재 확률을 고려한 MMSE(Minimum Mean Square Error) STSA(Short-Time Spectral Amplitude Estimator)를 전처리기로 도입하여 HMM(Hidden Markov Model)에 기반한 음성인식시스템의 인식성능을 평가하였다. 음성인식 시스템의 실시간 구현을 고려하여, MMSE STSA 기법을 음성개선을 위한 전처리기로 사용할 때 MMSE STSA의 이득계산 과정에서 많은 계산량이 요구되는 modified Bessel 함수를 근사화하여 사용하였다.

Ⅰ. 서 론

음성인식시스템의 인식성능을 저하시키는 음성신호의 왜곡은 크게 부가잡음에 의한 왜곡과 채널특성에 의한 왜곡으로 나눌 수 있다. 부가잡음은 음성신호에 합의 형태로 나타나고, 채널특성에 의한 왜곡은 음성신호에 대해 컨벌루션의 형태로 나타난다[1]. 때문에, 부가잡음에 의해 왜곡된 음성신호의 개선은 음성인식 시스템의 전처리 과정으로 많이 연구되며, 잡음제거를 위해 MMSE STSA와 같이 주파수 영역에서[2,3], 채널 특성의 차이는 CMN(Cepstral Mean Normalization) 등과 같이 켑스트럼 영역에서 연구가 많이 이루어져 왔다[4].

본 연구에서는 다양한 부가잡음에 의한 인식시스템의 성능저하를 줄이고자 음성 부재확률을 도입한 MMSE STSA 음성개선 기법을 도입하였으며, 그 계산부하를 줄이기 위해 이득 계산과정에서의 modified Bessel 함수를 근사화하여 사용하였다.

본 논문의 구성은 다음과 같다. Ⅰ장 서론에 이어 Ⅱ장에서는 MMSE STSA음성개선기법 및 modified Bessel 함수의 근사화에 대해 설명하고, Ⅲ장에서는 실험환경 및 부가잡음에 의해 왜곡된 신호에 음성개선기법을 전처리기로 사용하였을 경우의 인식결과를 제시하고, 결과를 검토한다. 마지막으로 Ⅳ장에서 결론을 맺고 향후 연구방향을 제시한다.

Ⅱ. MMSE STSA을 적용한 음성개선 기법

부가잡음에 의해 오염된 신호는 다음과 같이 나타낼 수 있다.

$$y(t) = x(t) + d(t) \tag{1}$$

여기서, $x(t)$와 $d(t)$는 각각 잡음에 의해 오염되지

않은 원신호와 잡음 신호를 나타낸다. 식 (1)을 주파수 영역으로 옮겼을 때 각 주파수 성분은 식 (2)와 같이 나타낼 수 있다.

$$Y_k = X_k + D_k \tag{2}$$

여기서, 첨자 k는 신호의 k번째 주파수 성분임을 나타낸다. 또한, 식 (2)의 각 주파수 성분은 아래와 같이 크기 및 위상 성분으로 나타낼 수 있다.

$$\begin{aligned} X_k &= A_k\, e^{j\alpha_k} \\ Y_k &= R_k\, e^{j\theta_k} \end{aligned} \tag{3}$$

인간의 청각 특성은 신호의 위상보다 진폭에 더욱 민감하므로 원 신호의 추정은 진폭 A_k를 추정하는 문제로 생각할 수 있다. 신호의 각 주파수 성분들이 독립적이라고 가정하면, A_k의 최소 자승 에러 추정은 식(4)로 구해진다[2].

$$\begin{aligned} \widehat{A_k} &= E\{A_k | y(t), 0 \le t \le T\} \\ &= E\{A_k | Y_0, Y_1, \ldots\} \\ &= E\{A_k | Y_k\} \\ &= \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} M(v_k) R_k \\ &= G_{MMSE}(\xi_k, \gamma_k) R_k \end{aligned} \tag{4}$$

$$\text{,where } v_k = \frac{\xi_k}{1+\xi_k}\gamma_k$$

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)}$$

$$\gamma_k = \frac{R_k^2}{\lambda_d(k)}$$

$$M(\theta) = e^{-\frac{\theta}{2}}\left[(1+\theta) I_0\left(\frac{\theta}{2}\right) + \theta I_1\left(\frac{\theta}{2}\right)\right]$$

여기서, $\lambda_x(k)$와 $\lambda_d(k)$는 각각 k번째 주파수 성분에서의 원신호와 잡음신호의 파워 추정치이다. $\Gamma(\cdot)$는 gamma 함수를 나타내며, ξ_k는 prio:i SNR, γ_k는 posterior SNR을 나타낸다. G_{MMSE}는 잡음신호에서 원신호를 추정하기 위한 이득이며, I_0, I_1은 각각 영차와 일차의 modified Bessel 함수를 나타낸다.

관측된 잡음신호에서 음성신호가 존재할 확률을 도입하여 음성개선에 도움을 줄 수 있는 것으로 알려져 있으며[3], 이를 위해 식 (4)에 음성부재확률을 도입하면 식 (5)와 같이 나타낼 수 있다.

$$\widehat{A_k} = \frac{\Lambda(\xi_k, \gamma_k, q_k)}{1+\Lambda(\xi_k, \gamma_k, q_k)} G_{MMSE}(\xi_k, \gamma_k) R_k \tag{5}$$

여기서, $\Lambda(\cdot)$는 다음과 같이 정의되는 generalized likelihood ratio이다.

$$\Lambda(Y_k, q_k) = \mu_k \frac{P(Y_k | H_k^1)}{P(Y_k | H_k^0)} \tag{6}$$

$$\text{,where } \mu_k = \frac{(1-q_k)}{q_k}$$

식 (6)에서 q_k는 k번째 주파수 성분이 존재하지 않을 확률이고, H_k^1, H_k^0는 각각 음성 존재와 부존재의 상태를 나타낸다.

식 (4)의 이득 계산부분에서 사용되는 modified Bessel 함수를 그림 1에 나타내었다. 실험에서는 [0, 10]의 v_k에 대해서는 영차와 일차의 변형 베셀함수에 대한 테이블을 만들어 사용하였으며, 그 이외의 구간에 대해서는 이득이로 식 (7)을 사용하였다.

$$G_{MMSE} = \frac{\xi_k}{1+\xi_k} \tag{7}$$

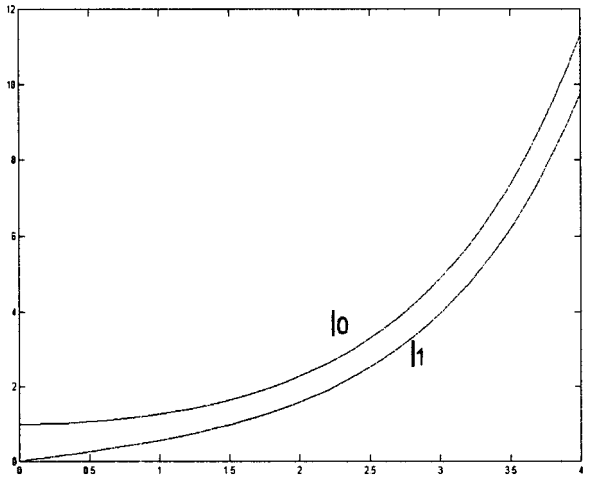

그림 1. Modified bessel functions of zero/first order

Ⅲ. 실험 및 검토

실험환경은 다음과 같다. MMSE STSA 음성개선기법의 적용에 따른 인식률의 변화를 보기 위해 PTM(Phonetic Tied-Mixture) 모델을 사용하여 가변어휘 인식기를 구현하였다. 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 [6].

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. 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 켑스트럼을 구하고, 현재 구간을 포함한 전후 각 6 구간(전체 13 구간)의 정보를 이용하여 1차의 차분 에너지와 12차의 차분 멜 켑스트럼, 그리고 차분-차분 값을 구하였다. 표 3-1에 음성의 분석 조건을 나타내었다.

휴대폰 환경으로의 적응을 위해 기법으로는 MAP (Maximum A Posteriori)를 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 3-2에 음성개선 기법을 적용하지 않았을 경우의 인식률을 나타내었으며, 표 3-3에 음성개선기법을 적용하였을 경우의 인식률을 나타내었다. 결과에서 볼 수 있듯이 휴대환경에서 modified Bessel 함수를 근사화를 적용한 MMSE STSA 음성개선 기법을 적용하였을 경우 전체적으로 인식률이 향상됨을 볼 수 있다.

표. 3-1 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

Ⅳ. 결 론

본 연구에서는 다양한 부가잡음에 의한 인식시스템의 성능저하를 줄이고자 음성 부재확률을 적용한 MMSE STSA 음성개선 기법을 전처리기로 도입하고, 그 계산부하를 줄이기 위해 이득 계산과정에서의 modified Bessel 함수의 테이블을 사용하여 근사화하였다.

인식실험 결과 휴대환경에서 modified Bessel 함수 근사화를 적용한 MMSE STSA 음성개선 기법이 적용되었을 경우 보다 높은 인식률을 얻을 수 있었다. 향후, 다양한 환경에서 강인한 음성인식을 위해 음성개선 기법과 특징추출부와의 통합과 채널 잡음에 대한 처리 기법등이 연구되어야 한다.

표 3-2. 기본 시스템에서의 인식률

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

표 3-3. 음성개선 기법을 적용한 시스템에서의 인식률

	Base	MAP	MAP+LDA
HHP SCH-600	90.16	93.37	89.16
HHP SPH-7000	94	91.61	93.7
HF SPH-7000	73.75	85.39	84.31
Total	84.44	88.94	89.01

본 연구는 한국과학재단 목적기초연구(R01-1999-00233)지원으로 수행되었음.

참 고 문 헌

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

[2] Yariv Ephraim and David Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[3] Liu Zhibin, Xu Naiping, "Speech Enhancement Based on Minimum Mean-Square Error Short-Time Spectral Estimation and Its Realization," *IEEE International Conference on Intelligent Processing Systems*, pp. 1794-1797, 1997.

[4] 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.

[5] D.L. Donoho, "De-Noising by Soft-Tresholding," *IEEE Trans. on Information Theory*, pp. 961-1005, 1995.

[6] 손종목, 정성윤, 배건성, "휴대폰 단말기에 적용을 위한 강인한 음성인식," *신호처리합동학술대회 논문지*, pp. 495-498, 2001.