신뢰성 높은 서브밴드 선택을 이용한 잡음에 강인한 화자식별

김 성 탁, 지 미 경, 김 회 린
한국정보통신대학교 공학부

Noise Robust Speaker Identification using Reliable Sub-Band Selection in Multi-Band Approach

Sungtak Kim, Mikyong Ji, and Hoirin Kim
School of Engineering, Information and Communications University
E-mail : stkim@icu.ac.kr, lindaji@icu.ac.kr, hrkim@icu.ac.kr

Abstract

The conventional feature recombination technique is very effective in the band-limited noise condition, but in broad-band noise condition, the conventional feature recombination technique does not produce notable performance improvement compared with the full-band system. To cope with this drawback, we introduce a new technique of sub-band likelihood computation in the feature recombination, and propose a new feature recombination method by using this sub-band likelihood computation. Furthermore, the reliable sub-band selection based on the signal-to-noise ratio is used to improve the performance of this proposed feature recombination. Experimental results shows that the average error reduction rate in various noise condition is more than 27% compared with the conventional full-band speaker identification system.

I. 서론

화자식별은 화자의 음성신호를 이용하여 등록된 화자들 중에서 가장 유사한 화자를 찾아내는 것이다. 최근에는 가우시안 혼합 모델 (GMM)을 이용한 문맥 독립 화자식별 기술[1]이 주된 추세이다. 화자모델링을 위한 특징벡터로는 멜프리퀀시켑스트럼 계수 (MFCC)를 많이 사용한다. 기존의 멜프리퀀시켑스럼 계수를 구하는 방법은 전체 주파수 밴드를 이용한다. 전체주파수를 이용하여 특징벡터를 구하는 방법의 경우, 음성신호가 비록 주파수 영역이 제한된 잡음 (Band-limited noise)으로 왜곡이 되더라도 전체 특징벡터 성분에 영향을 주게 된다. 이런 문제점을 극복하기 위해 다중밴드 (Multi-band) 방법이 제안되었다. 다중밴드방법에는 크게 유사도 재조합 (Likelihood recombination)과 특징벡터 재조합 (Feature recombination)방법으로 나누어진다. 하지만, 기존의 다중밴드 방법은 음성신호가 광대역 노이즈 (Broad-band noise)에 의해 왜곡된 경우에는 성능향상에 크게 기여를 못한다.

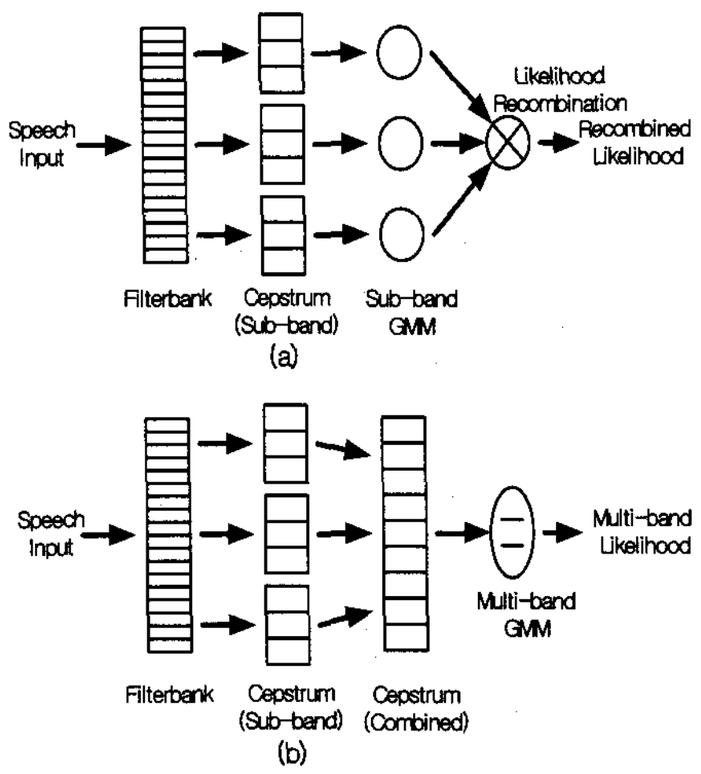

그림 1. 다중밴드 방법 (a) 유사도 재조합 방법 (b) 특징벡터 재조합 방법

하지만, 음성신호가 광대역 노이즈에 의해 왜곡이 되었더라도 각 서브밴드들의 왜곡정도는 다르다. 이런 점을 감안하면, 다중밴드 방법이 광대역 잡음환경에서도 유용한 방법임을 알 수 있다. 기존의 특징벡터 재조합방법은 유사도를 구하는 과정에서 전체 서브밴드 벡터들을 모두 사용해서 구한다. 이렇게 구한 유사도는 각 서브밴드들의 왜곡정도를 반영하지 못한다. 본 논문에선 기존의 방법의 단점을 극복하기 위해 서브밴드 유사도 계산법을 소개하고, 이 서브밴드 유사도 계산법을 이용하여 변형된 특징벡터 재조합 방법을 제안한다. 더 나아가 서브밴드 신호 대 잡음비를 이용하여 신뢰성 높은 서브밴드 선택 방법을 이용하여 제안된 변형된 특징벡터 재조합 방법의 성능을 더욱 향상 시켰다.

II. 다중밴드 MFCC[2]를 이용한 서브밴드 유사도 계산

M 개의 서브밴드와 N개의 필터를 가지는 시스템에서 각 서브밴드 당 L개의 MFCC를 추출한다면, i번째 서브밴드의 j번째 MFCC를 구하는 방법은 식 (1)과 같다.

$$x_j^{(i)} = \sqrt{2/NM} \sum_{n=1}^{N/M} LFB_n^{(i)} \cos\left((n-0.5)\frac{j\pi}{N/M}\right), 1 \le j \le L \le \frac{N}{M} \quad (1)$$

여기서 $LFB_n^{(i)}$는 i번째 서브밴드의 n번째 필터에너지의 로그 값이다. 그림 2는 두 개의 서브밴드와 N개의 필터를 가지는 멀티밴드 시스템에서의 멀티밴드 MFCC 추출방법을 보여준다.

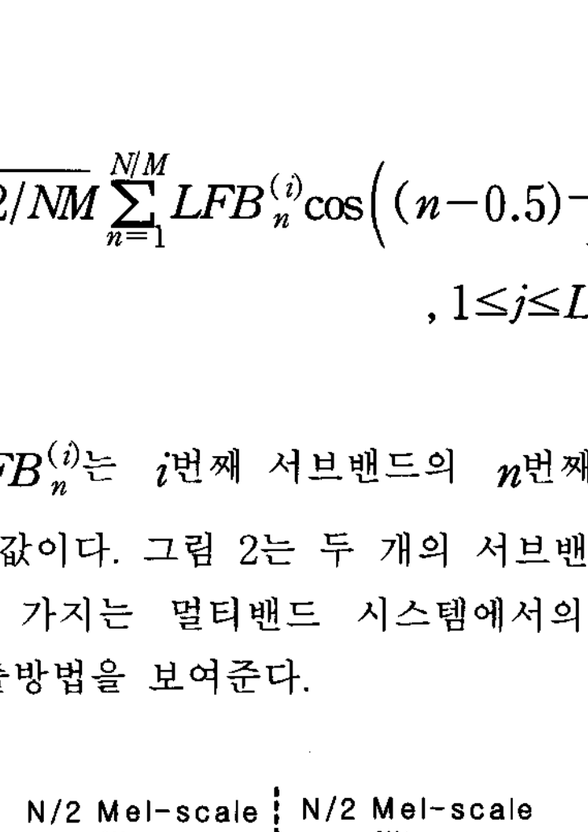

그림 2. 두 개의 서브밴드를 가지는 멀티밴드 시스템에서 MFCC 추출방법

M개의 서브밴드를 가지는 멀티밴드 시스템에서 재조합된 특징벡터 X는 각 서브밴드 특징벡터로 나누어진다. 즉 $X=\{x^1, x^2, \ldots, x^M\}$. 그리고, 각 서브밴드 특징벡터들이 확률적으로 독립이라 가정하면, 각 서브밴드들의 유사도 값들을 구할 수 있다.

$$p(X|\lambda) = \sum_{w=1}^{W} C_{w,\lambda} p(x^1, x^2, \ldots, x^M|w,\lambda) = \sum_{w=1}^{W} C_{w,\lambda} \prod_{i=1}^{M} p(x^i|w,\lambda) \quad (2)$$

식 (2)에서 구하고자하는 서브밴드를 제외한 나머지 밴드들에 대해 marginalization을 하면 구할 수 있게 된다.

$$= \sum_{w=1}^{W} C_{w,\lambda} p(x^i|w,\lambda) \prod_{m=1, m\neq i}^{M} \int p(x^m|w,\lambda) dx^m \frac{p(x^i|\lambda)}{} \quad (3)$$
$$= \sum_{w=1}^{W} C_{w,\lambda} p(x^i|w,\lambda)$$

그림 3은 서브밴드 유사도를 이용한 특징벡터 재조합 방법을 그림으로 보여준다. 그림 3을 보면 그림 1의 기존의 특징벡터 재조합 방법과 다르다는 것을 알 수 있다.

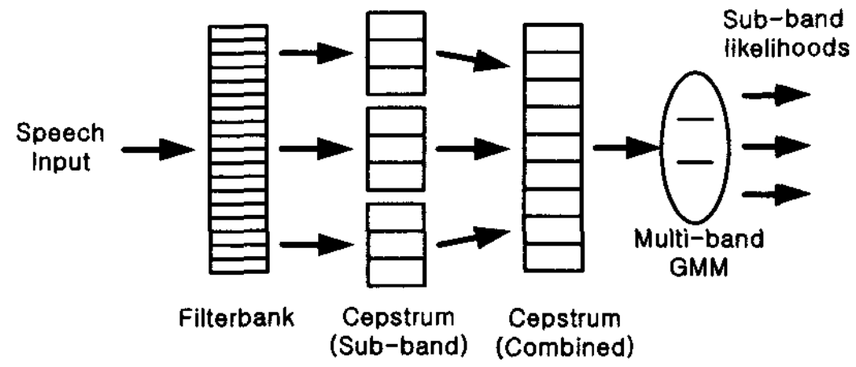

그림 3. 서브밴드 유사도를 이용한 특징벡터 재조합 방법

III. 신뢰성 높은 서브밴드 선택을 이용한 화자식별

본 논문에선 신뢰성 높은 서브밴드를 선택하기 위해 신호 대 잡음비를 이용하였다. 노이즈 에너지는 비 음성 프레임들의 평균 에너지를 사용하였다. 각 프레임을 음성과 비 음성 프레임으로 판별하는 방법은 처음 10개의 프레임들의 평균 에너지보다 크면 음성 프레임으로, 작으면 비 음성 프레임으로 간주하였다. t번째 프레임의 신호 대 잡음비는 식 (4)와 같다.

$$SNR_t^{Full} = 10\log_{10}\left[\frac{\sum_{k=1}^{K}|S_t(k)|^2}{\sum_{k=1}^{K}|\overline{N}(k)|^2}\right] \quad (4)$$

$$|S_t(k)| = \max\{|X_t(k)| - 1.1|\overline{N}(k)|, 0.001|\overline{N}(k)|\} \quad (5)$$

서브밴드 시스템에서 t번째 프레임의 i번째 서브밴드 신호 대 잡음비는 식 (5)를 이용하여 구할 수 있다.

$$SNR_t^i = 10\log_{10}\left[\frac{\sum_{k\in Sub-band\ i}|S_t(k)|^2}{\sum_{k\in Sub-band\ i}|\overline{N}(k)|^2}\right] \quad (6)$$

여기서 k, $|X_t(k)|$, $|S_t(k)|$,그리고 $|\overline{N}(k)|$는 각각 주파수 인덱스, 잡음 신호의 에너지 절대값, 추정된 무 잡음 신호의 에너지 절대값, 그리고 노이즈의 에너지 절대값을 나타낸다. 식 (2), (3), 그리고 (6)을 이용해서 서브밴드 유사도와 신뢰성 높은 서브밴드 선택을 이용한 화자식별은 아래 식 (7)과 같다.

$$\hat{S} = \arg\max_{S}\frac{1}{N_{rb}}\sum_{t=1}^{T}\left(\sum_{i=1}^{M}\log\left(p(x_t^i|\lambda_k)\right)\right) \quad (7)$$

$$\log\left(p(x_t^i|\lambda_k)\right) = \begin{cases}\log\left(p(x_t^i|\lambda_k)\right) & \text{if } SNR_t^i \geq threshold \\ 0 & otherwise\end{cases} \quad (8)$$

λ_k는 k번째 화자모델을 나타내고, N_{rb}는 신뢰성 높은 서브밴드의 수를 나타낸다. 신뢰성 높은 프레임 선택을 이용한 기존의 전체밴드를 이용한 방법과 기존의 특징벡터 재조합 방법을 사용한 화자식별은 식 (9)와 같다. 여기서 N_{rf}는 신뢰성 높은 프레임의 수를 나타낸다.

$$\hat{S} = \arg\max_{S}\frac{1}{N_{rf}}\sum_{t=1}^{T}\log(p(X_t|\lambda_k)) \quad (9)$$

$$\log(p(X_t|\lambda_k)) = \begin{cases}\log(p(X_t|\lambda_k)) & \text{if } SNR_t^{Full} \geq threshold \\ 0 & otherwise\end{cases} \quad (10)$$

논문에서는 임계값을 30dB로 정하였다.

IV. 실험 및 결과

본 논문에선 실험을 위해 TIMIT 데이터베이스를 사용하였다. 등록화자는 여자 100명과 남자 100명으로 총 200명을 사용하였다. 화자모델은 Maximum A Posteriori (MAP) 알고리즘을 이용하여 모델링하였다 [3]. 이때, 사용된 Universal Background Model (UBM)을 추정하기 위해 남자 50명과 여자 50명을 사용하였다. 등록 화자 당 10문장을 발성하였고 5문장은 화자모델링에 나머지 5문장은 테스트에 사용하였다. 노이즈 환경을 위해 16kHz를 8kHz로 다운샘플링을 하고 Aurora 2 데이터베이스[4]의 8가지 잡음 (Airport, Babble, Car, Exhibition, Restaurant, Street, Subway, Train)을 여러 가지 신호 대 잡음비로 음성을 왜곡시켰다. 실험에 사용한 화자모델과 UBM은 160개의 mixture를 가지는 Gaussian mixture model을 사용 하였다. 기존의 전체 주파수 밴드를 이용한 방법에선 33개의 필터를 가지는 필터뱅크를 이용하여 18차의 MFCC를 추출하였다. 다중밴드를 이용한 특징벡터 재조합 방법의 필터뱅크의 필터 수와 MFCC의 차수를 표 1에 나타내었다.

표 1. 특징벡터 재결합 방법의 필터 수와 MFCC 차수

System / Parameters	Multi-band system		
	2 sub-bands	3 sub-bands	4 sub-bands
N	32	33	32
L(Total)	9(18)	6(18)	4(16)

표 2. 전체밴드를 이용한 방법 (FULL_BAND)의 화자식별 결과 (%)와 2개의 서브밴드 (CFR_2), 3개의 서브밴드 (CFR_3), 그리고 4개의 서브밴드 (CFR_4)를 가지는 기존의 특징벡터 재조합 방법들의 에러감소율(%)

Method / SBRs	FULL-BAND Accuracy (%)	Error Reduction Rate (%)		
		CFR_2	CFR_3	CFR_4
20dB	81.64	-2.25	2.18	-0.41
15dB	67.10	7.18	8.40	5.47
10dB	47.10	5.60	5.27	3.57
5dB	28.13	0.77	0.82	-0.45
Average ERR		2.82	4.15	2.04

표 2는 기존의 전체 밴드를 이용한 방법과 다중밴드를 이용한 특징벡터 재결합 방법의 노이즈 환경에서의 화자식별 결과를 보여준다. 표 2의 결과를 보면 기존의 특징벡터 재결합 방법의 경우 광대역 노이즈 환경에서

큰 성능 향상을 보여주지 못 한다. 표 3은 기존의 3개의 서브밴드를 가지는 특징벡터 재결합 방법과 서브밴드 유사도를 이용한 변형된 특징벡터 재조합 방법의 화자식별 결과를 보여준다.

표 3. 기존의 특징벡터 재결합 방법 (CRF_3)과 서브밴드 유사도를 이용한 변형된 특징벡터 재결합 방법 (FRS_3)의 전체밴드를 이용한 방법에 대한 에러 감소율

Method / SNRs	Error Reduction Rate (%) over full-band	
	CRF_3	FRS_3
20dB	2.18	17.29
15dB	8.40	16.11
10dB	5.27	8.13
5dB	0.82	1.06

표 3의 결과를 보면 서브밴드 유사도를 이용한 변형된 특징벡터 재결합 방법이 기존의 방법보다 특히 높은 SNR에서 좋은 성능을 보여준다. 표 4는 신뢰성 높은 프레임 선택을 이용한 특징벡터 재조합방법과 신뢰성 높은 서브밴드 선택을 이용한 변형된 특징벡터 재조합 방법의 성능을 나타내었다.

표 4. 신뢰성 높은 프레임 선택을 이용한 특징벡터 재조합방법 (CFR+RFS)과 신뢰성 높은 서브밴드 선택을 이용한 변형된 특징벡터 재조합방법 (FRS+RSS)을 이용한 화자식별 결과

Method / SNRs	Error Reduction Rate (%) over full-band	
	CFR+RFS	FRS+RSS
20dB	4.97	32.33
15dB	15.39	37.08
10dB	15.19	28.31
5dB	7.63	10.66
Ave.	10.80	27.10

표4의 결과를 보면 신뢰성 높은 서브밴드 선택을 이용한 변형된 특징벡터 재조합방법이 신뢰성 높은 프레임 선택을 이용한 기존의 특징벡터 재조합방법에 비해 좋은 성능을 보여주었다.

V. 결론

본 논문에서는 기존의 특징벡터 재조합의 단점인 서브밴드 전체를 이용하여 유사도를 구하는 방법을 극복하기 위해 서브벤드 유사도 계산법을 소개하고, 서브밴드 유사도를 이용한 변형된 특징벡터 재조합 방법을 제안하였다. 더 나아가 서브밴드 신호 대 잡음비를 이용해 신뢰성 높은 서브밴드 선택방법을 이용해 잡음에 강인한 화자식별 알고리즘을 제안하고 여러 가지 잡음과 SNR에 대해 실험을 하였다. 그 결과, 제안된 방법을 사용한 화자식별 결과가 기존의 방법들보다 우수한 성능을 보여주었다.

참고문헌

[1] D. Reynold and R. C. Rose, "Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models," *Proc. IEEE Trans. Speech and Audio Processing*, Vol. 3, pp. 72-83, Jan. 1995.

[2] B. Mak, "A Mathematical Relationship Between Full-Band and Multiband Mel-Frequency Cepstral Coefficients," *IEEE Signal Processing Letters*, Vol. 9(8), pp. 241-244, 2002.

[3] D. Reynold, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, Vol. 10, pp. 19-41, 2000.

[4] D. Pearce and H. Hirsch, "The Aurora Experimental Framework for The Performance Evaluation of Speech Recognition under Noise Conditions," in Proc. ICSLP, Vol. 4, pp. 29-32, 2000.