말소리와 음성과학 제6권 제1호(2014.03.31)
ISSN 2005-8063, pp. 77~83
www.speechsciences.or.kr
http://dx.doi.org/10.13064/KSSS.2014.6.1.077

수정된 MAP 적응 기법을 이용한 음성 데이터 자동 군집화

Automatic Clustering of Speech Data Using Modified MAP Adaptation Technique

반 성 민[1] · 강 병 옥[2] · 김 형 순[3]
Ban, Sung Min · Kang, Byung Ok · Kim, Hyung Soon

ABSTRACT

This paper proposes a speaker and environment clustering method in order to overcome the degradation of the speech recognition performance caused by various noise and speaker characteristics. In this paper, instead of using the distance between Gaussian mixture model (GMM) weight vectors as in the Google's approach, the distance between the adapted mean vectors based on the modified maximum a posteriori (MAP) adaptation is used as a distance measure for vector quantization (VQ) clustering. According to our experiments on the simulation data generated by adding noise to clean speech, the proposed clustering method yields error rate reduction of 10.6% compared with baseline speaker-independent (SI) model, which is slightly better performance than the Google's approach.

Keywords: speech recognition, speech data clustering, KL divergence, MAP adaptation

1. 서론

음성인식 시스템의 성능을 저하시키는 대표적인 요인으로 훈련 환경과 테스트 환경의 불일치 문제를 들 수 있으며, 특히 화자 및 환경 특성 등의 차이로 인한 음향 변이가 이러한 불일치의 핵심 요인이다. 이러한 음향 변이를 음성 특징 벡터 및 모델 영역에서 보상하고자 하는 많은 시도들이 제안되었으며, 그 예로 히스토그램 등화, 벡터 테일러 급수, 병렬 모델 조합 등이 있다[1]-[3]. 이러한 보상 방식들이 음향 환경 불일치 문제를 어느 정도 완화시켜 주기는 하지만 성능 향상의 한계를 보인다. 그 외에도 화자 및 환경 적응 방식을 통해 이러한 음향 변이 문제를 해결하려는 시도들이 있는데[4]-[6] 실제 응용 분야에 따라서는 테스트 음성과 동일한 화자 및 환경의 적응 데이터를 확보하기 어려운 경우가 많이 있다. 음향 변이에 대처하는 가장 간단한 방법은 매우 많은 수의 화자 및 환경 특성을 포함하는 대용량의 음성 데이터베이스를 구축하여 이로부터 다양한 음향 변이를 포함하는 음향 모델을 구성하는 방법이다. 최근 스마트폰을 이용한 음성 검색 서비스 등이 활성화됨에 따라 지속적으로 대용량의 음성 데이터 수집이 가능하게 되었고, 이를 음향 모델 훈련에 적용함으로써 음성인식 성능이 많이 향상되었다. 이처럼 다양한 음향 변이를 포함하는 대용량 음성 데이터베이스의 확보가 가능한 경우, 단일 음향 모델을 구성하는 것보다는 다양한 음향 특성 그룹 별로 복수의 음향 모델들을 구성함으로써 추가적인 인식성능 향상을 기대할 수 있다.

본 논문에서는 유사한 특성을 가지는 복수의 음향 모델을 구성하기 위해 대용량의 음성 데이터베이스를 자동 군집화한다. 이와 같은 음성 데이터 군집화에 관한 연구가 일부 수행되어 왔는데, 연구 초기에는 사전에 음성 데이터 별로 제공된 환경 정보를 이용하여 군집화를 하는 방법이 제안되었고, 최근에는 사전 정보 없이 데이터를 자동 군집화하는 방법들이 제안되었다[7]-[9]. 이 중 구글에서는 가우스 혼합 가중치에 대한 Kullback-Leibler (KL) 발산을 이용한 군집화 방식을 제안

1) 부산대학교, bansungmin@pusan.ac.kr
2) 한국전자통신연구원, bokang@etri.re.kr
3) 부산대학교, kimhs@pusan.ac.kr, 교신저자

이 논문은 미래창조과학부의 ETRI 연구개발지원사업의 일환으로 수행되었으며(지원번호: 11921-03001, "Beyond 스마트 TV 기술개발"), 논문의 일부는 2012년 한국음성학회 가을 학술대회에서 발표된 바 있다[12].

접수일자: 2014년 1월 7일
수정일자: 2014년 2월 24일
게재결정: 2014년 3월 15일

하였고[7], 마이크로소프트에서는 i-vector를 이용한 방식을 제안하였다[8]. 본 논문에서는 maximum a posteriori (MAP) 적응 기법을 이용하여 적응된 평균 벡터 기반의 군집화 방식을 제안하고, 기존의 구글 방식과 성능을 비교한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서 기존의 구글 군집화 방식을 설명하고, 3장에서 제안한 방식에 대해 설명한다. 4장에서는 다양한 군집화 방식의 성능을 비교하고 5장에서 논문의 결론을 맺는다.

2. 구글의 군집화 방식

앞에서 언급한 것처럼 구글에서는 화자와 환경 정보를 이용하지 않고도 음성 데이터를 자동 군집화하는 기술을 개발하였다[7]. 군집화는 벡터양자화(Vector Quantization (VQ))를 이용하여 수행하고, 군집과 입력 음성 사이의 거리 척도로는 아래 식과 같이 가우스 혼합 모델(Gaussian Mixture Model (GMM)) 사이의 KL 발산을 사용한다.

$$D(p^U|p^C) = \int_{\boldsymbol{o}} p^U(\boldsymbol{o})\log\frac{p^U(\boldsymbol{o})}{p^C(\boldsymbol{o})} = \int_{\boldsymbol{o}} [\sum_{j=1}^{M} w_j^U g_j(\boldsymbol{o})]\log[\sum_{l=1}^{M} w_l^U g_l(\boldsymbol{o}) / \sum_{l=1}^{M} w_l^C g_l(\boldsymbol{o})] \quad (1)$$

여기서 p^U와 p^C는 입력 음성과 군집에 속한 데이터를 사용하여 적응한 모델의 확률 분포를 나타낸다. $g_j(\boldsymbol{o})$는 전체 훈련 DB로부터 구한 음향모델의 j번째 가우스 분포에 대한 관측 벡터 $\boldsymbol{o}$의 확률을 나타내고, w_j^U와 w_j^C는 j번째 가우스 분포에서의 적응된 가중치를 나타낸다. 계산적인 효율성을 위해 음향모델에서 평균과 분산은 제외 하고, 가중치만을 적응한다. 가중치의 적응을 위해 아래와 같이 expectation maximization (EM) 알고리즘을 수행한다.

$$w_j^U = \frac{1}{K}\sum_{k=1}^{K}\frac{w_j g_j(\boldsymbol{u}_k)}{\sum_{l=1}^{M} w_l g_l(\boldsymbol{u}_k)} \quad (2)$$

여기서 w_j는 전체 훈련 DB로부터 구한 음향모델의 j번째 가우스 분포에 대한 가중치이다. $\boldsymbol{u}_k$는 입력 음성 U가 포함하는 특정 상태에 대한 관측벡터 K개 중 k번째 관측벡터를 나타낸다. 식 (1)에서 j번째 가우스 분포가 $w_j^U g_j(\boldsymbol{o}) \gg 0$로 주요한 경우에 아래 식과 같이 근사화가 가능하다.

$$w_j^U g_j(\boldsymbol{o})\log[\sum_{l=1}^{M} w_l^U g_l(\boldsymbol{o})] \cong w_j^U g_j(\boldsymbol{o})\log[w_j^U g_j(\boldsymbol{o})] \quad (3)$$

식 (3)을 식 (1)에 대입하면

$$D(p^U|p^C) \cong \sum_{j=1}^{M} w_j^U \log\frac{w_j^U}{w_j^C} = D(W^U|W^C) \quad (4)$$

인데, 여기서 $W^U = w_1^U,\cdots,w_M^U$, $W^C = w_1^C,\cdots,w_M^C$ 이고, 각각은 입력 음성 U와 군집으로부터 적응된 모델의 M개 가중치를 나타낸다. 입력 음성 U가 포함하고 있는 어휘에 독립적으로 군집화 하기 위해 U가 포함하는 모든 상태에 대한 KL 발산은

$$D_{Total}(p^U|p^C) = \frac{1}{N^U}\sum_i D(p_i^U|p_i^C) \quad (5)$$

와 같이 정규화시킨다. 여기서 N^U는 입력 음성 U의 정렬 결과에서 관찰된 GMM의 수이고, 입력 음성과 군집 사이의 KL 발산 $D(p_i^U|p_i^C)$를 GMM의 개수 N^U로 정규화한다. 이 때 p_i^U와 p_i^C는 i번째 GMM에서 입력 음성 U와 군집 데이터 C를 사용하여 적응한 모델들을 나타낸다. 모든 군집과 입력 음성 사이의 $D_{Total}(p^U|p^C)$를 이용하여 입력 음성과 가장 가까운 군집을 구하고, VQ 결과가 수렴할 때까지 반복한다. VQ 과정에서 초기 군집은 임의로 정하거나, 주성분 분석을 이용하여 훈련 DB를 2개 군집으로 분할한다.

3. MAP 적응을 이용한 군집화 방식

본 논문에서는 구글의 군집화 방식을 기반으로 하되, 음소의 상태별 GMM 가중치 벡터 사이의 거리를 사용하는 대신에 상태별로 적응 평균 벡터 사이의 거리를 사용하는 방식을 제안한다. 이를 위해 MAP 적응 방식을 사용하여 상태별 적응 평균 벡터를 구한다. MAP 적응 방식을 적용한 적응 평균 벡터는 아래와 같이 구할 수 있다. 편의상 상태에 대한 인덱스는 생략한다.

$$\boldsymbol{\mu}_m^U = \alpha\boldsymbol{\mu}_m^{SI} + (1-\alpha)\boldsymbol{\mu}^U \quad (6)$$

$$\boldsymbol{\mu}_m^C = \alpha\boldsymbol{\mu}_m^{SI} + (1-\alpha)\boldsymbol{\mu}^C \quad (7)$$

여기서 $\boldsymbol{\mu}_m^U$와 $\boldsymbol{\mu}_m^C$는 각각 입력 발화와 군집 데이터를 적응 데이터로 했을 때 m번째 혼합의 적응 평균 벡터이고, $\boldsymbol{\mu}^U$와 $\boldsymbol{\mu}^C$는 각각 입력 발화와 군집 데이터에 대한 표본평균 벡터를 나타낸다. $\boldsymbol{\mu}_m^{SI}$는 화자 독립(Speaker Independent (SI)) 모델의 m번째 혼합 벡터이고, α는 $\boldsymbol{\mu}_m^{SI}$와 표본평균 사이의 가중치이다. 음성인식에서의 MAP 적응은 안정적인 인식 성능을 확보하기

위해서 α를 적응 데이터의 프레임 수에 따라 결정한다[10]. 하지만 이렇게 구한 적응 평균 벡터를 군집화에 이용하면 환경에 따라 군집화 되는 대신에, 관측 벡터의 길이에 따라 군집화 되는 문제가 발생할 수 있다. 이러한 문제를 해결하기 위해 본 논문에서는 기존의 MAP 적응 방식을 수정하여 가중치 α를 상수로 둔다. 군집과의 거리측정은 구글 군집화 방식과 동일하게 KL 발산을 이용하지만, GMM 사이의 KL 발산에 대한 닫힌 해가 존재하지 않기 때문에, KL 발산의 상한치로 근사화하여 이용한다[11].

$$D(p^U,p^C) \cong \sum_{m=1}^{M}\sum_{d=1}^{D} w_m \frac{(\mu_{m,d}^U-\mu_{m,d}^C)^2}{(\sigma_{m,d}^C)^2} \tag{8}$$

여기서 p^U와 p^C는 식 (1)과 마찬가지로 각각 입력 문장과 군집 데이터로부터 SI 음향모델을 적응시킨 음소의 상태별 확률분포를 나타낸다. 또한 D는 특징 벡터의 차원을 나타내고, $\mu_{m,d}^U$와 $\mu_{m,d}^C$는 각각 적응 평균 벡터 $\boldsymbol{\mu}_m^U$와 $\boldsymbol{\mu}_m^C$의 d번째 차원의 값을 나타낸다. 식 (8)에서 차원별 분산 값들이 일정하다고 가정하면(편의상 1이라고 둠) 다음과 같이 간략화 된다.

$$D(p^U,p^C) \cong \sum_{m=1}^{M}\sum_{d=1}^{D} w_m (\mu_{m,d}^U-\mu_{m,d}^C)^2 \tag{9}$$

식 (9)를 벡터에 관한 식으로 바꾼 후, 식 (6), (7)을 대입하면

$$\begin{aligned} D(p^U,p^C) &\cong \sum_{m=1}^{M} w_m |\boldsymbol{\mu}_m^U-\boldsymbol{\mu}_m^C|^2 \\ &= (1-\alpha)|(\boldsymbol{\mu}^U-\boldsymbol{\mu}^C)|^2 \end{aligned} \tag{10}$$

이 되고, 여기서 상수 부분인 $1-\alpha$을 생략하면

$$\tilde{D}(p^U,p^C) = |\boldsymbol{\mu}^U-\boldsymbol{\mu}^C|^2 \tag{11}$$

와 같이 표본평균에 대한 유클리드 거리로 나타낼 수 있고, 이는 참고문헌 [12]의 식과 동일하다.

상태별로 적응된 가중치 벡터와 평균 벡터가 데이터 군집화에 미치는 영향은 다를 수 있다. 가중치 벡터는 식 (2)와 같이 입력 데이터가 베이스라인 음향모델의 GMM 분포에 부합하는 상대적인 비율을 나타낸다. 이러한 특성도 음성 데이터별 환경 차이를 어느 정도 반영하지만, 상태별로 적응된 평균 벡터가 음성 데이터별 환경차이를 좀 더 직접적으로 반영할 수 있을 것으로 기대한다.

4. 실험 및 결과

4.1 실험 환경

제안한 군집화 방식과 기존의 군집화 방식의 성능을 평가하기 위해서 다양한 잡음을 깨끗한(clean) 음성에 인위적으로 더하여 시뮬레이션 음성 DB를 구성하였다. 깨끗한 음성 DB로는 훈련용으로 ETRI에서 제작한 POW DB를 사용하였고 [13], 테스트용으로 SiTEC에서 제작한 PBW 452 DB를 사용하였다[14]. 잡음은 Aurora 2 DB 평가셋에 포함된 8종류의 잡음(airport, babble, street, train, car, restaurant, subway, exhibition)을 사용하였다. 이 중 일부(airport, babble, street, train)는 훈련 DB 구성에 사용하였고, 나머지(car, restaurant, subway, exhibition)는 테스트 DB 구성에 사용하였다. 또한 훈련 및 테스트 DB의 SNR은 5dB, 10dB, 15dB, 20dB로 하였다. 전체 훈련 DB의 길이는 묵음 구간을 포함하여 약 12시간이고 총 발화 수는 61,000 발화이다.

음성인식을 위한 음향모델은 트라이폰 단위의 HMM을 사용하고, 트리 기반 군집화를 이용하여 상태 tying된 3 상태의 left-to-right HMM을 사용한다. SI 모델과 각 군집별 모델에서 사용하는 가우스 혼합은 11개를 사용하였다. 음성인식기는 HTK를 사용하였고, 특징벡터는 위너 필터를 적용한 후[15], 39차 MFCC를 추출하였다. 군집별 음향모델 중 테스트 데이터에 부합하는 최적의 모델은 GMM 우도가 최대가 되도록 하는 음향모델로 선정하였다.

일반적으로 훈련 데이터의 수가 증가할수록 정교한 음향모델을 얻을 수 있어서 음성인식 성능도 증가하지만 훈련 데이터의 수가 어느 수준 이상 증가하면 음성인식 성능은 수렴하여 더 이상 증가하지 않는다. 전체 훈련 데이터의 수 N_{total}에서 음성인식 성능이 수렴한 경우에 훈련 데이터를 L개로 군집화하여 각각의 음향 모델을 훈련할 때, 각 군집에 속한 데이터 수는 평균적으로 N_{total}/L로 감소하게 된다. 따라서 각 군집별로 감소된 데이터로 훈련한 음향모델은 성능 저하의 요인이 될 수 있다. 하지만 전체 훈련 데이터의 수가 LN_{total} 이상이면 L개 군집 각각의 훈련 데이터만으로도 충분히 정교한 군집별 음향모델을 훈련할 수 있다. 실제로 많은 경우 이처럼 음향모델이 충분히 수렴할 정도의 데이터를 가지고 있지 않기 때문에, 본 논문에서는 군집별 데이터를 적응 데이터로 이용하여 베이스라인 음향 모델을 적응한 모델을 사용한다. 적응 과정은 maximum likelihood linear regression (MLLR) 적응을 수행한 후 MAP 적응 방식을 적용한다.

<표 1>은 구글의 군집화 방식에서 노드 수가 7개일 때 각 군집별 음향 모델을 군집 데이터를 이용하여 재훈련한 경우(+retraining)와 군집 데이터를 적응 데이터로 이용하여 적응한 경우(+adaptation)의 단어오류율(Word Error Rate (WER))을 비

교한 것이다. 앞서 언급한 것처럼 전체 훈련 데이터가 충분하지 않은 경우, 감소된 군집별 데이터로 인한 성능 저하는 적응 과정을 통해 베이스라인 대비 8.54%의 오류감소율(Error Reduction Rate (ERR))로 성능 개선이 이루어졌음을 확인할 수 있다.

표 1. 음향모델 적응에 따른 구글 방식의 성능평가
Table 1. Performance evaluation of Google's clustering method according to adaptation technique

Method	WER (%)	ERR (%)
Baseline	10.63	·
Google (+retraining)	12.85	-20.96
Google (+adaptation)	9.72	8.54

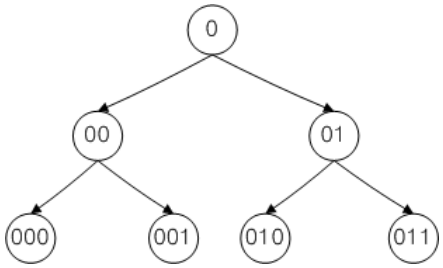

그림 1. 군집 노드의 트리구조
Figure 1. Tree structure of cluster node

4.2 실험 결과

구글의 군집화 방식을 시뮬레이션 DB에 적용하고, <그림 1>과 같은 트리구조의 7개 군집을 사용했을 때의 인식결과는 <표 2>와 같다. 각 군집별로 단어오류율과 오류감소율을 구하였고, 군집화 결과의 양상을 살펴보기 위해서 각 군집별 데이터의 분포를 <그림 2>에 나타내었다. 구글의 군집화 방식에서는 군집별 성능이 모든 군집들에서 베이스라인 성능보다 우수하다. 전체적으로 베이스라인 성능에 비해서 약 8.56%의 오류감소율이 얻어졌다. <그림 2>의 군집별 데이터 분포특성에서 노드 0은 SNR에 따라 노드 00과 노드 01로 분할되는 경향이 있는데, 노드 00에는 SNR이 높은 데이터가 많이 분포하며, 노드 01에는 SNR이 낮은 데이터가 많이 분포한다. 노드 00은 성별에 따라 노드 000과 노드 001로 분할되는 경향이 있는데, 노드 000은 남성의 데이터가 많이 분포하며, 노드 001은 여성의 데이터가 많이 분포하는데, SNR이 높을수록 분할이 잘 된다. 노드 01은 SNR이 낮은 데이터들로 군집화 되어 있어서 노드 00에 비해서 성별에 따른 분할이 상대적으로 잘 이루어지지 않고, 오히려 SNR에 따른 분할에 더 가깝다. 노드 010은 SNR이 높은 데이터가 많이 분포하며, 노드 011은 SNR이 낮은 데이터가 많이 분포한다.

표 2. 구글 군집화 방식에서 군집별 음성인식 성능
Table 2. Speech recognition performance for each cluster in the Google's clustering method

Node	Data size	WER (%)		ERR (%)
		Baseline	Clustering	
0	576	9.37	9.37	·
00	191	13.61	12.57	7.69
01	1096	9.12	8.12	11.00
000	268	6.72	5.97	11.11
001	2596	9.55	9.21	3.63
010	2291	5.59	4.85	13.28
011	2022	19.12	17.09	10.62
Total	9040	10.63	9.72	8.56

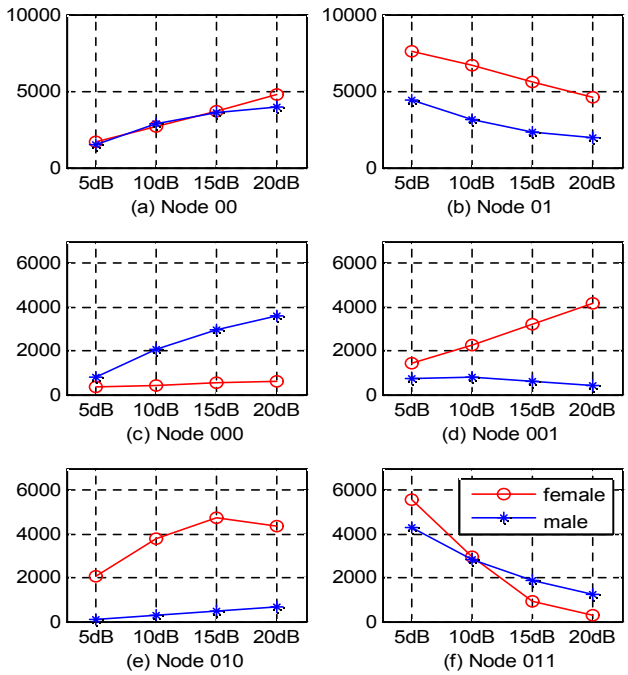

그림 2. 구글의 군집화 방식 사용 시 군집별 데이터 분포
Figure 2. Data distribution of each cluster in case of the Google's clustering method

<표 3>은 제안한 적응 평균 기반의 군집화 방식을 사용했을 때의 음성인식 성능을 나타낸다. 적응 평균을 이용한 군집화의 경우, 베이스라인 대비 10.63%의 오류감소율로 성능 개선이 이루어졌음을 확인할 수 있다. 이는 구글의 방식보다 오류감소율의 측면에서 2.07% 정도 우수하다. <그림 3>의 군집별 데이터 분포 특성을 살펴보면 노드 0에서의 SNR에 따른 분할이 구글 방식보다 잘 이루어진 것을 관찰할 수 있다. 또한 노드 00에서의 성별에 따른 분할도 아주 잘 이루어져 있다. SNR이 낮은 데이터들이 주로 분포하고 있는 노드 01은 성별에 따라 분할되지 않았지만 노드 011은 낮은 SNR의 데이터들 위주로 군집화 되어 있다.

표 3. 제안한 군집화 방식에서 군집별 음성인식 성능
Table 3. Speech recognition performance for each cluster in the proposed clustering method

Node	Data size	WER (%)		ERR (%)
		Baseline	Clustering	
0	637	10.99	10.99	·
00	896	3.46	4.02	-16.13
01	576	13.72	12.67	7.59
000	1244	3.30	3.46	-4.88
001	2619	4.54	4.09	10.08
010	640	7.66	8.12	-6.12
011	2428	23.60	19.73	16.40
Total	9040	10.63	9.50	10.63

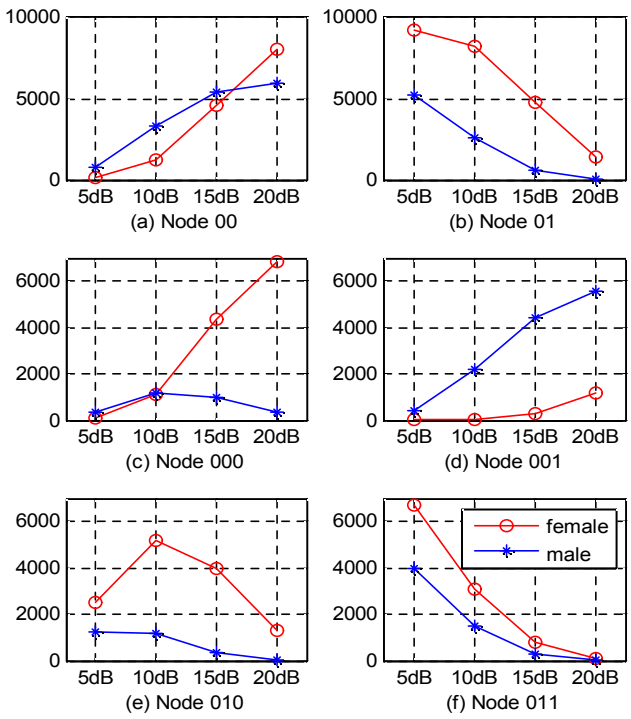

그림 3. 제안된 군집화 방식 사용 시 군집별 데이터 분포
Figure 3. Data distribution of each cluster in case of the proposed clustering method

<그림 2>와 <그림 3>을 비교했을 때 제안한 방식은 구글방식에 비해서 SNR 및 성별에 따른 군집화가 잘 이루어지는 것을 확인할 수 있다. 이는 <표 2>와 <표 3>에서 각 군집별 베이스라인 인식 성능의 편차 정도를 통해서도 확인할 수 있다. 군집화가 잘 될수록 이를 이용한 음성인식 성능이 향상될 수 있지만, 군집 선정이 정확하지 않다면 성능향상을 보장할 수 없다. 일부 군집 선정에 오류가 있는 경우, 제안한 군집화 방식 보다는 구글 방식의 군집들이 좀 더 다양한 환경 특성을 포함하고 있어서 군집 선정 오류에 더 강인할 수 있다. <표 2>와 <표 3>에서 구글 방식과 달리 제안한 방식에서는 일부 노드의 성능이 베이스라인보다도 못한 것을 관찰할 수 있는데, 이는 제안한 방식에서 군집화로 인한 성능 향상 보다 군집 선정 오류로 인한 성능 저하가 더 크기 때문으로 추정된다. 가장 열악한 데이터들이 모여 있는 노드 011의 경우에는 실제로 군집 선정 오류가 적었고, 베이스라인 성능에 대한 오류감소율이 16.40% 로 성능 향상 비율이 가장 높았다.

표 4. SNR 기반 군집화 방식에서 군집별 음성인식 성능
Table 4. Speech recognition performance for each cluster in the SNR based clustering method

Node	Data size	WER (%)		ERR (%)
		Baseline	Clustering	
0	560	9.46	9.46	·
00	1159	13.37	13.55	-1.29
01	990	3.54	3.64	-2.86
000	1927	25.95	25.12	3.20
001	925	10.16	9.73	4.26
010	981	6.32	7.14	-12.90
011	2498	2.48	2.80	-12.90
Total	9040	10.63	10.62	0.10

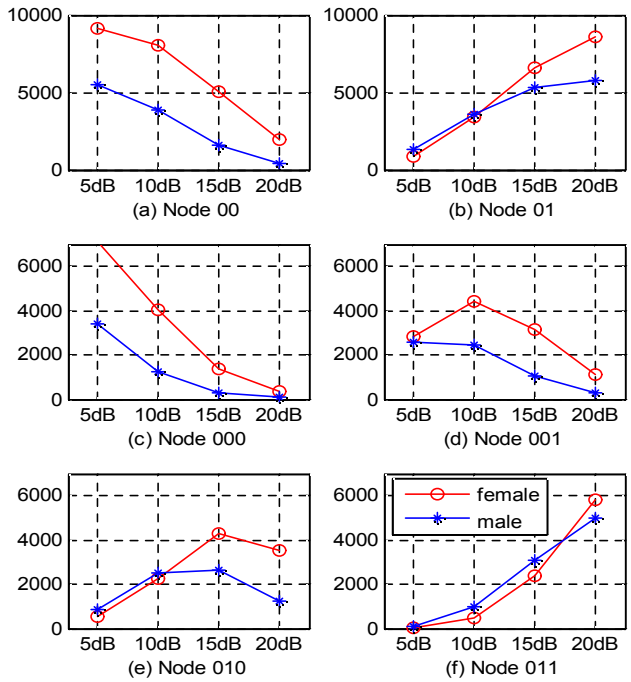

그림 4. SNR 기반 군집화 방식 사용 시 군집별 데이터 분포
Figure 4. Data distribution of each cluster in case of the SNR based clustering method

<표 4>는 SNR을 군집화에 사용했을 때의 음성인식 성능을 나타내고, <그림 4>는 SNR에 따른 군집별 데이터 분포를 나타낸다. SNR은 위너 필터링 된 신호의 음성구간과 묵음구간을 검출하여 음성에 대한 잡음의 에너지 비율로 추정하였다.

<그림 4>에서 군집별 데이터 분포를 보면 SNR별로 데이터들이 군집화 되어 있는 것을 확인할 수 있다. 모든 데이터들은 잡음 제거 모듈을 통과한 후에 SNR을 구하기 때문에 군집화 오류들이 존재한다. SNR을 군집화에 적용했을 때 인식 성능은 베이스라인에 비해 성능향상이 거의 없었고, 일부 노드에서는 베이스라인보다도 못한 성능을 보였다. 이는 다양한 종류의 잡음이 부가된 음성 특성을 단순히 SNR에 의거하여 군집화하는 것이 인식성능에 실제로 도움되지 않음을 의미한다.

<표 5>는 군집 노드 수에 따른 방식별 전체 인식 성능을 나타낸다. 군집화를 했을 때 베이스라인 성능에 비해서 오류감소율 향상이 있는 것을 확인할 수 있고, SNR 기반 군집화를 제외한 모든 방식들에서 노드 수는 3개를 사용하는 것보다 7개를 사용했을 때 성능이 더 좋았다. 그리고 본 논문에서 제안한 적응 평균 방식이 약 10.6%의 오류감소율을 얻어 가장 성능이 우수하였다.

표 5. 군집 노드 수에 따른 군집화 방식 별 음성인식 성능
Table 5. Speech recognition performances of clustering methods according to the number of cluster nodes

Clustering method	WER (%)		ERR (%)
	3 node	7 node	
Baseline	10.63	10.63	-
Google (adapted weight)	10.15	**9.72**	8.56
Proposed (adapted mean)	9.85	**9.50**	10.63
SNR	**10.61**	10.62	0.19

5. 결론

본 논문에서는 화자 및 환경 특성 차이로 인한 음성인식 성능 저하를 극복하기 위해 수정된 MAP 적응 방식을 사용하여 음소의 상태별 적응 평균 벡터 사이의 유클리드 거리를 이용하는 군집화 방식을 제안하고, 그 군집화 성능을 검토하였다. 시뮬레이션 데이터를 이용한 실험 결과에서 제안한 군집화 방식은 SI 모델과 비교하여 10.6%의 음성인식 오류감소율을 보였으며, 기존 구글의 군집화 방식에 비해 조금 개선된 성능을 보였다.

참고문헌

[1] Hilgerk, F., Molau S., & Ney H. (2002). Quantile based histogram equalization for online applications. *Proc. ICSLP,* 237-240.

[2] Moreno, P. J., Raj B., & Stern, R. M. (1996). A vector Taylor series approach for environment-independent speech recognition. *Proc. ICASSP*, 733-736.

[3] Gales, M. J. F. & Young, S. J. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Trans. on Speech and Audio Process*, 5(5), 352-359.

[4] Deng, L., Droppo, J., & Acero A. (2003). Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Trans. Speech Audio Process*, 11, 6, 568-580.

[5] Gales, M. J. F. (1997). Maximum likelihood linear transformations for HMM based speech recognition. *Cambridge Univ. Tech. Rep. TR 291*, Cambridge, U.K.

[6] Song, H. J., Jeon, H. B. & Kim, H. S. (2009). Fast speaker adaptation based on eigenspace-based MLLR using artificially distorted speech in car noise environment. *Phonetics and Speech Sciences*, 1(4), 119-125.
(송화전, 전형배, 김형순 (2009). 차량 잡음 환경에서 인위적 왜곡 음성을 이용한 Eigenspace-based MLLR에 기반한 고속 화자 적응, 말소리와 음성과학, 1(4), 119-125.)

[7] Beaufays, F., Vanhoucke, V., & Strope, B. (2010). Unsupervised discovery and training of maximally dissimilar cluster models. *Proc. Interspeech*, 66-69.

[8] Zhang, Y., Xu, J., Yan, Z. J., & Huo, Q. (2011). An i-vector based approach to training data clustering for improved speech recognition. *Proc. Interspeech*, 1247-1250.

[9] Tsao, Y. & Lee, C. H. (2009). An ensemble speaker and speaking environment modeling approach to robust speech recognition. *IEEE Trans. Audio, Speech, and Language Processing*, 17(5), 1025-1037.

[10] Lee, C. H., Lin, C. H. & Juang, B. H. (1991). A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Transactions on Signal Processing*, 39(4), 806-814.

[11] Campbell, W. M., Sturim, D. E., Reynolds, D. A. & Solomonoff, A. (2006). SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. *Proc. ICASSP*, 1, 97-100.

[12] Ban, S. M., Kang, B. O., Lee, Y. K., & Kim, H. S. (2012). Automatic clustering of speech data using the distance between the cepstral mean vectors. *Proc. 2012 Fall Conf. of the Korean Society of Speech Sciences*, 35-36.
(반성민, 강병옥, 이윤근, 김형순 (2012). 켑스트럼 평균벡터 거리를 이용한 음성 데이터 자동 클러스터링, 한국음성학회 가을 학술대회 발표논문집, 35-36.)

[13] Lim, Y. & Lee Y. (1995). Implementation of the POW (phonetically optimized words) algorithm for speech database. *Proc. ICASSP*, 1, 89-92.

[14] Lee, Y. J., Kim, B. W., Kim, J. J., Yang, O. Y. & Lim, S. Y. (1995). Some considerations for construction of PBW set. *Proc. 12th Workshop on Speech Communications and Signal Processing*. Korean Association of Speech Sciences, 310-314. (이용주, 김봉완, 김종진, 양옥렬, 임선영 (1995). 음성 DB용 PBW에 관한 검토, 제12회 음성통신 및 신호처리 워크샵 논문집, 한국음성학회, 310-314.)

[15] Lee, S. J., Kang, B. O., Jung, H. Y., Lee, Y. K. & Kim, H. S. (2010). Statistical model-based noise reduction approach for car interior applications to speech recognition. *ETRI Journal,* 32(5), 801-809.

• **반성민 (Ban, Sung Min)**
부산대학교 전자전기컴퓨터공학부
부산시 금정구 장전2동 부산대학로 63번길
Tel: 051-510-1704 Fax: 051-510-4279
Email: bansungmin@pusan.ac.kr
관심분야: 음성인식, 음성 전처리
현재 전자전기컴퓨터공학부 대학원 박사과정 재학 중

• **강병옥 (Kang, Byung Ok)**
한국전자통신연구원 음성처리연구실
대전시 유성구 가정동 가정로 218
Tel: 042-860-5684 Fax: 042-860-4889
Email: bokang@etri.re.kr

• **김형순 (Kim, Hyung Soon) 교신저자**
부산대학교 전자전기컴퓨터공학부
부산시 금정구 장전2동 부산대학로 63번길
Tel: 051-510-2452
Email: kimhs@pusan.ac.kr
관심분야: 음성인식, 음성합성, 음성신호처리
1992~현재 전자전기컴퓨터공학부 교수