한국음향학회지 제22권 제5호 pp. 383~387 (2003)

화자검증을 위한 새로운 코호트 선택 방법

A New Method of Selecting Cohort for Speaker Verification

김 성 준*, 계 영 철*
(Seong-Jun Kim*, Young-Chul Kay*)

*홍익대학교 전자공학과

(접수일자: 2003년 2월 11일; 수정일자: 2003년 6월 26일; 채택일자: 2003년 6월 30일)

본 논문에서는 기존의 고정크기의 코호트 집단을 기반으로 한 화자검증 방법을 다룬다. 특히, 본 논문에서는 고정크기의 코호트 대신에 화자모델들 사이의 거리를 이용하는 가변크기의 새로운 코호트를 제안한다: 제안된 새로운 방식에서는 각 화자로부터 일정한 거리 내에 있는 주변 화자모델들의 밀집도가 고려된다. 그 화자 주변의 밀집도가 높으면 코호트의 크기가 자동적으로 증가되어 화자검증률이 개선되고, 반면 밀집도가 적으면 코호트의 크기가 감소되어 계산량이 줄어든다. 실험결과 제안된 방법이 기존의 방식에 비하여 EER (Equal Error Rate)을 감소시킴을 확인할 수 있었다.

핵심용어: 화자검증, 코호트, 스코어 정규화, 동적 임계값
투고분야: 음성처리 분야 (2.5)

This paper deals with the method of speaker verification based on the conventional cohort of fixed size. In particular, a new cohort of variable size, which makes use of the distance between speaker models, is proposed: The density of neighboring speaker models within the fixed distance from each speaker is taken into account in the proposed method. The high density leads to the increase of cohort size, thus improving the speaker verification rate. On the other hand, the low density leads to its decrease, thus reducing the amount of computations. The simulation results show that the proposed method outperforms the conventional one, achieving a reduction in the EER.

Keywords: *Speaker verification, Cohort, Score normalization, Dynamic threshold*
ASK subject classification: *Speech signal processing (2.5)*

I. 서 론

최근 보안 분야가 발전함에 따라서 개인의 음성 특성을 이용한 화자검증 분야의 연구가 꾸준히 이루어져 왔다. 화자검증의 가장 간단한 방법으로는 입력으로 들어온 음성의 스코어 (score)와 해당되는 화자의 임계값을 비교하여 수락이나 거절을 결정하는 것이다. 그러나 같은 화자일지라도 발음의 변동이 있거나 발음환경이 변하면 음성의 스코어가 변하기 때문에, 화자마다 고정된 임계값을 사용하는 방식은 인식률의 저하를 가져온다. 따라서 환경의 변화에 따라 같이 변화되는 동적 (dynamic) 임계값을 사용하는 방법이 제안되었다[1].

특히, 각각의 화자에 대하여 그와 유사한 음성특성을 갖는 일정한 수의 화자들로 구성되는 코호트 (cohort) 집단을 이용하여 동적 임계값을 결정하는 방식이 제안되었다[1,2,4]. 그러나 이러한 기존의 방식은 화자의 특성에 관계없이 모두 동일한 크기의 코호트를 사용하므로 다른 화자와 구별이 잘 안되는 화자의 경우에는 코호트의 크기[1]가 부족하여 화자 검증률이 저하되는 문제가 발생한다. 또한 구별이 쉬운 화자의 경우는 적은 정보량으로도 화자검증이 가능한데 비하여 많은 정보량을 사용함으로써 계산량이 증가하는 단점이 있다.

본 논문에서는 화자의 특성을 고려하여 화자마다 다른 크기의 코호트를 사용하는 방법 (즉, 서로 구별이 어려운 화자의 경우에는 코호트 크기를 증가시키고 반대의 경우는 감소시키는 방법)을 제안한다. 이를 위하여 각 화자사이의 거리를 계산하여 각 화자로부터 일정한 거리 내에 있는 화자들을 그 화자의 코호트 집단에 속하도록 한다. 이렇게 함으로써 발음한 화자가 다른 화자와 구별되기가

책임저자: 계영철 (yckay@wow.hongik.ac.kr)
서울시 마포구 상수동 72-1
홍익대학교 전자공학과
(전화: 02-320-1604; 팩스: 02-320-1119)

1) 코호트의 크기는 코호트 집단에 속하는 구성원의 수로 정의한다.

어려우면 (즉, 발음한 화자와 비슷한 화자가 많으면) 많은 수의 화자가 코호트에 속하게 되고, 그렇지 않으면 반대의 경우가 된다.

본 논문에서는 일정한 거리 내에 있는 화자의 수와 실험으로 정해진 하한경계 (lower limit) 중 더 큰 값을 코호트의 최종크기로 선택하여 인식률 실험을 하고 기존의 코호트 방식과 성능의 비교분석을 하였다.

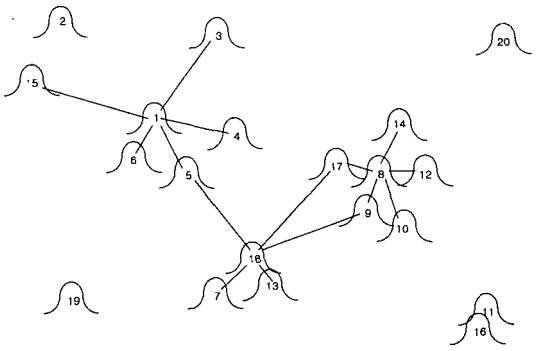

그림 1. 고정크기의 코호트의 선택
Fig. 1. The selection of a cohort of fixed size.

II. 본 론

2.1. 화자간의 거리

각 화자들에 대한 코호트 집단을 선택하기 위해서 먼저 화자들간의 거리를 구해야 한다. i번째 화자와 j번째 화자 사이의 거리는 다음과 같이 정의한다[1].

$$D(i,\ j) = \left\{ \sum_{k=1}^{M} \log P(O_{ik}|\lambda_i) + \sum_{k=1}^{M} \log P(O_{ik}|\lambda_j) \right\} / 2M \quad (1)$$

여기서 O_{ik}는 화자 i의 k번째 관측된 벡터이고, λ_i는 화자 i의 HMM (Hidden Markov Model) 모델이며 M은 학습시 필요한 관측된 벡터의 개수이다. 화자 i와 화자 j 사이의 거리는 식 (1)과 같이 관측된 벡터가 화자 i에 의해서 발음되었을 때 모델 j와의 유사도와 관측된 벡터가 화자 j에 의해서 발음되었을 때 모델 i와의 유사도의 평균으로 정의된다.

2.2. 고정 크기의 코호트 집단에 의한 화자검증

일반적으로 화자검증을 하기 위해서는 식 (2)과 같은 판별식을 이용한다.

$$P(O|\lambda_{s=I}) \underset{\text{false}}{\overset{\text{true}}{\gtrless}} T(s=I) \quad (2)$$

여기서 I는 발음한 화자의 신원 (ID)을 의미하고 $P(O|\lambda_{s=I})$는 입력음성 O가 화자 I에 의해 발음되었을 유사도 (likelihood)이며, $T(s=I)$는 화자 I에 속한 임계값이다. 즉 $P(O|\lambda_{s=I})$를 측정하여 화자 I의 임계값 $T(s=I)$ 보다 크게 되면 수락되고 그렇지 않으면 거절되게 된다. 식 (2)에서 알 수 있듯이 임계값 $T(s=I)$를 어떻게 결정을 하느냐에 따라서 화자 검증률에서 차이가 나게 된다.

임계값을 결정하는 효과적인 방법 중의 하나는 화자 I를 제외한 화자들 중에서 화자 I와 비슷한 발음을 가진 K명의 화자들을 화자 I의 코호트 집단으로 선택한 후 그것들로부터 임계값을 결정하는 것이다[1]. 즉, 화자들 간의 거리 식 (1)을 계산한 후에 각 화자와 가장 가까운 K개의 집합을 그 화자의 코호트 집단으로 선택하고 동적 (dynamic) 임계값을 식 (3)과 같이 결정한다[1,2,4].

$$T(s=I) = \max_{1 \le k \le K} P(O|C_k(\lambda_{s=I})) \quad (3)$$

여기에서 $C_k(\lambda_{s=I})$는 화자 I의 코호트 집단의 k번째 구성원을 나타낸다. 이렇게 하면 화자의 목소리 변이, 채널 부정합 (mismatch) 등의 문제로 인해 $P(O|\lambda_{s=I})$의 값이 낮게 나올 경우 동적 임계값 $P(O|C_k(\lambda_{s=I}))$의 값도 따라서 낮게 나오므로 화자검증의 오류를 줄일 수 있다.

그림 1은 고정된 크기의 코호트 선택을 설명하는 것으로 각 모델들 사이의 거리에 관계없이 K=5 명의 화자를 코호트 모델로 선택하는 것을 나타낸다.

2.3. 제안된 가변 크기의 코호트 집단

기존의 고정 크기의 코호트 모델을 사용하면 화자 I와 유사한 화자의 수가 코호트 크기보다 클 경우 코호트 크기의 부족으로 인하여 화자 검증률이 저하된다. 또한 그 반대의 경우에는 작은 크기의 코호트를 할당하여도 정확한 검증이 가능한데 오히려 큰 크기의 코호트를 할당하였으므로 계산량이 증가되는 단점이 있다. 즉, 화자 I와 비슷한 화자가 많을 경우에는 $\lambda_{s=I}$ 주위로 다른 모델들이 밀집해 있는 경우로 이 때는 사칭자 (imposter)를 구별해 내기 어려운 상황이 된다. 그러므로 이러한 경우에는 코호트 집단의 크기를 어느 정도 크게 늘려주어야 화자검증

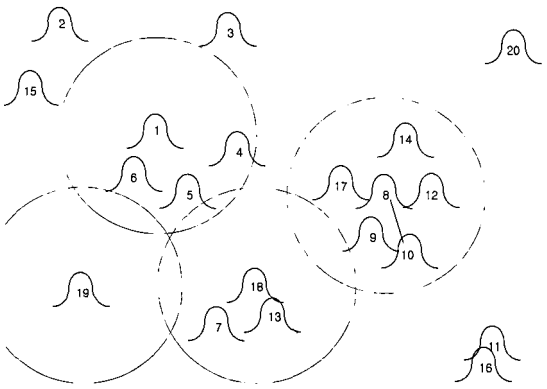

그림 2. 고정거리를 이용한 코호트의 선택
Fig. 2. The selection of a cohort utilizing fixed distance.

률이 개선된다. 또한 화자 I와 비슷한 화자가 적을 경우에는 $\lambda_{s=I}$ 주위로 다른 모델들이 넓게 퍼져있는 경우로 이때는 사칭자와 쉽게 구별해 낼 수 있다. 그러므로 이러한 경우는 적은 크기의 코호트 집단만을 선택하여도 비슷한 화자검증률을 낼 수 있으므로 적은 코호트 집단을 선택하여 계산량을 줄일 수 있게 된다.

이러한 점에 착안하여 본 논문에서는 $\lambda_{s=I}$ 로부터 일정한 거리 내에 있는 모델들을 모두 코호트 집단으로 선택하는 방법을 제안한다. 즉 일정거리 내에 $s\neq I$ 모델들이 많은 경우는 해당되는 화자와 비슷한 발음을 하는 사람들이 많다는 것이므로 그 화자들을 모두 코호트 모델로 선택하게 되고, 반대로 일정한 거리 내에 있는 모델들이 적은 경우는 해당되는 화자와 비슷한 발음을 하는 사람들이 적다는 것이므로 다른 화자들과 쉽게 구분이 될 수 있으며 그만큼 코호트 크기를 줄여주어도 인식률이 저하되지 않는다.

그림 2는 일정한 거리를 경계로 그 안쪽에 있는 모델들을 모두 각 화자의 코호트 모델로 선택하는 것을 나타낸다. 그림 2를 살펴보면 19번 모델은 일정한 거리 내에 한 명의 화자도 포함하고 있지 않아 식 (3)을 구할 수가 없음을 알 수 있다.

이러한 문제는 코호트 집단의 크기에 하한한계 (lower limit)를 할당함으로써 해결가능하다. 즉 다음과 같이 최종적인 코호트 집단의 크기를 결정한다.

$$\text{코호트 크기} = \max(\text{거리를 이용한 코호트 크기}, \text{lower limit}) \quad (4)$$

이렇게 최종 코호트 크기를 결정을 하면 사칭자와 구별이 잘 안되는 화자의 경우에는 코호트 크기가 크게 되며, 사칭자와 구별이 잘 되는 경우는 코호트 크기를 너무 낮추지 않는 범위 한도 내까지, 즉 하한 한계까지 낮춤으로 해서 계산량을 줄여줄 수 있다.

III. 실 험

본 실험에 사용된 음성 데이터베이스는 홍익대학교 학생 80명이 4자리의 숫자 2개를 각각 40번씩 발음한 음성을 16 bit, 16 kHz 샘플링으로 녹음하여 구성되었다. 그 중에서 4번의 발음은 모델을 만들기 위한 학습 (training) 음성으로 사용하였으며, 나머지 발음은 화자검증 테스트를 위하여 사용하였다. HMM의 상태 수는 1자리 숫자당 3개의 상태를 사용하였고 특징벡터는 에너지와 12차 MFCC (Mel-Frequency Cepstral Coefficient) 벡터, Delta, Accelation 계수를 사용하여 총 39차 벡터를 사용하였다. 또한 6개의 가우시안 분포를 사용하여 연속적 HMM 모델을 사용하여 각 화자당 모델을 생성하였다. 먼저 80명의 화자에 대하여 각 화자의 4개의 음성을 이용하여 80개의 HMM 모델 $\lambda_1 \sim \lambda_{80}$ 을 만들고, 그때의 학습 음성과 모델 $\lambda_1 \sim \lambda_{80}$을 이용하여 각 화자들 사이의 거리 행렬 (distance matrix)을 구하였다[1]. 모델들 사이의 거리 행렬을 구한 후에, 그것을 이용하여 각 모델에 대하여 거리가 가까운 모델부터 리스트를 작성하였다. 이 리스트를 기초로 하여 고정크기의 코호트인 경우는 거리가 작은 모델부터 이미 정해진 수 K 만큼을 각 화자에 할당하고, 가변크기의 코호트인 경우는 정해진 거리 내에 들어오는 모든 모델들을 각 화자에 할당하여 사용하였다.

표 1은 고정크기의 코호트를 사용하였을 경우의 화자

표 1. 고정크기 코호트인 경우의 EER
Table 1. EER for the cohort of fixed size.

cohort size	EER (%)	cohort size	EER (%)
1	35.78	14	0.92
2	16.71	15	0.92
3	10.65	16	0.92
4	7.1	17	0.92
5	3.94	18	0.78
6	3.68	19	0.78
7	1.84	20	0.78
8	1.71	21	0.78
9	0.92	22	0.78
10	0.92	23	0.78
11	0.92	24	0.78
12	0.92	25	0.78
13	0.92		0.78

표 2. 고정거리 D=20 일 때 EER
Table 2. EER for fixed distance D=20.

1	4.8	0.92	14	14.38	0.92
2	5.16	0.92	15	15.31	0.92
3	5.66	0.92	16	16.25	0.92
4	6.26	0.92	17	17.2	0.92
5	6.9	0.92	18	18.16	0.78
6	7.55	0.92	19	19.13	0.78
7	8.3	0.92	20	20.1	0.78
8	9.08	0.92	21	21.06	0.78
9	9.88	0.92	22	22.05	0.78
10	10.73	0.92	23	23.03	0.78
11	11.63	0.92	24	24.01	0.78
12	12.53	0.92	25	25	0.78
13	13.45	0.92			

검증률을 EER (Equal Error Rate)로 나타낸 것이다. 표 1로부터 알 수 있듯이 코호트의 크기가 커질수록 EER은 줄어들게 되며 크기가 작은 쪽에서는 화자 검증률이 아주 좋지 않음을 알 수가 있다.

표 2는 거리 D=20 이내에 들어오는 모델들을 코호트로 사용한 경우이며, 앞서 언급한 식 (4) 하한경계를 변화시켜 가면서 측정한 EER을 나타내고 있다.

표 3은 각각의 거리에 따른 EER을 나타낸 것이다. 고정거리 내에 코호트가 하나도 포함되지 않는 경우를 방지하기 위하여 하한한계를 1로 취하였으며 식 (4), 거리를 변화시켜가면서 EER을 구하였다.

고정거리를 이용하는 코호트 방식에서는 코호트의 크기가 화자마다 다르게 되므로 여기에서는 평균크기를 사용하였다. 표 2와 3의 EER과 표 1의 EER을 같은 코호트 크기에 대하여 비교하여 보면 전자의 경우가 EER이 줄어듬을 알 수 있으며, 이는 표 2와 3에서 진하게 표시된 부분으로 나타내고 있다. 표 2를 살펴보면 하한경계를 충분히 증가시키는 경우 평균크기와 EER이 표 1의 그것에 접근함을 알 수 있다. 이는 하한경계값이 거리에 의하여 결정된 코호트의 크기보다 커져 결국 고정크기의 코호트와 동일하게 되기 때문이다 (식 4).

코호트의 거리 D를 여러 가지로 변화시키면서 실험한 결과에 의하면, 각 화자로부터 가장 가까이 있는 첫 번째 코호트 모델인 $C_1(\lambda_{s=I})$, I=1...80 까지의 거리들의 평균을 D로 정하고, 하한경계는 1~2개 정도를 사용하면 계산량을 줄이면서 화자검증률이 향상됨을 확인할 수 있었다.

표 3. 하한한계=1 때 거리에 따른 EER
Table 3. EER for various distances with lower limit=1.

15	1.36	35.78	23	10.36	0.92
16	1.63	35.78	24	13.23	0.92
17	2.16	10.65	25	16.46	0.78
18	2.81	7.1	26	20.43	0.78
19	3.58	1.71	27	24.28	0.78
20	4.8	0.92	28	28.15	0.78
21	6.11	0.92	29	32.3	0.78
22	8.18	0.92	30	35.56	0.78

IV. 결 론

본 논문에서는 동적 임계값을 이용하는 기존의 고정크기의 코호트 집단의 성능을 개선하는 방법을 제안하였다. 고정된 크기의 코호트를 사용하는 기존방식 대신에, 본 논문에서는 고정된 거리를 이용하는 가변크기의 코호트를 사용하는 방식을 제안하였다. 즉 각 화자의 주변 모델들의 밀집도를 이용하여 그 화자 주변의 밀집도가 높으면 코호트의 크기를 증가시켜 화자 검증률을 개선시키고, 밀집도가 적으면 코호트의 크기를 감소시켜 계산량을 줄일수 있도록 하였다.

실험결과 제안된 방법의 EER이 기존 방식의 EER보다 감소됨을 확인하였다.

참 고 문 헌

1. A. E. Rosenberg, J. Delong, C.-H. Lee, B.-H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker verification," *Proc. ICSLP 92*, 1, 599-602, 1992.
2. H. Cho, S. Ku, and K. Shi, "A new cohort normalization using local acoustic information for speaker verification" *Proc. ICASSP '99*, 2, 841-844, 1999.
3. R. A. Finan, A. T. Sapeluk, and R. I. Damper, "Imposter cohort selection for score normalisation in speaker verification," *Journal of Pattern Recognition Letter*, 18, 881-888, 1997.
4. C. S. Liu, H. C. Wang, and C.-H. Lee, "Speaker verification using normalized log-likelihood score," *IEEE Trans. on Speech and Audio Processing*, 4 (1), 56-60, 1996.
5. L. Rabiner, and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall International Inc, 321-389, 1993.

저자 약력

• 김 성 준 (Seong-Jun Kim)

1998년 2월: 동신대학교 전기전자공학과 학사
2000년 2월: 홍익대학교 전자공학과 석사
2000년 2월~현재: 홍익대학교 전자공학과 박사 과정
※ 주관심분야: 음성 및 영상인식, 화자인식, 디지털 신호처리

• 계 영 철 (Young-Chul Kay)

1980년 2월: 서울대학교 전자공학과 학사
1982년 2월: 한국과학기술원 전기 및 전자공학과 석사
1991년 5월: Univ. of Southern California, Electrical Eng. Ph.D.
1991년 9월~현재: 홍익대학교 전자전기공학부 부교수
※ 주관심분야: 디지털 신호처리, 음성 및 영상인식, 로봇 비젼