

문장종속형 화자확인에서의 관측확률 가중기법

김세현[○] 장길진 오영환

한국과학기술원 전산학과

Observation Probability Weighting Method for Text-Dependent Speaker Verification

Se-Hyun Kim[○] Gil-Jin Jang Yung-Hwan Oh

Department of Computer Science

Korea Advanced Institute of Science and Technology

{shkim,jangbal}@bulsai.kaist.ac.kr, yhoh@cs.kaist.ac.kr

요 약

기존의 문장종속형 화자인식 방법들은 대부분 음성인식에서 사용되는 방법을 그대로 적용하기 때문에, 화자의 개인성 정보보다 음운정보에 더 민감한 단점이 있다. 화자인식 시스템의 성능향상을 위해서는 음운정보보다는 화자의 개인성 정보가 잘 반영되도록 하는 것이 중요하다. 본 논문에서는 HMM(hidden Markov model)을 기반으로 한 문장종속형 화자확인 시스템의 성능향상을 위한 관측확률 가중방법을 제안한다. 먼저 주어진 학습자료에서 화자의 개인성이 잘 반영된 프레임들을 예측한다. 임의의 입력음성에 대한 인식점수는 화자의 특징이 잘 반영된 프레임의 관측확률에 가중치를 주어 구한다. 제안한 방법을 적용한 결과 기존의 우도비(likelihood ratio) 정규화 점수를 사용하는 방법에 비해 동일오류율(EER; equal error rate)을 2~3%정도 줄여 인식을 향상시킬 수 있었다.

1 서론

화자인식은 화자가 발성한 음성을 이용하여 발성한 화자에 대한 정보를 알아내는 기술로서, 기존의 보안시스템인 지문인식이나 홍채인식 등과는 달리 원격지에서 전화 등을 이용하여 인식을 수행할 수 있다는 장점을 가지고 있다. 화자인식은 출력 결과에 따라서 크게 화자 확인(speaker verification)과 화자 식별(speaker identification)로 나누어진다[4]. 화자확인 은 발성된 음성이 원하는 화자인지 아닌지를 구분해 내는 것이며, 화자식별은 등록된 여러 명의 사람들 중 입력 음성에 가장 가까운 화자를 찾아내는 것을 말한다. 또 다른 분류로는 인식을 위해 발성해야 하는 문장의 형태에 따라 문장종속형(text dependent)과 문장자유형(text independent)으로 나누어진다.

기존의 문장종속형 화자확인에서 사용되는 방법들은 음성 인식에서와 별 차이가 없어 화자의 개인성 정보를 충분히 반

영하지 못하였다. 음성인식에서는 시스템 적응을 위하여 화자간의 개인차를 극복해야 하지만, 화자인식의 경우 이러한 개인차를 인식에 이용해야만 좋은 성능을 얻을 수 있다. 제안된 방법은 주어진 학습자료를 이용하여 화자의 개인성 정보가 잘 반영되어 있는 프레임들을 선택한 후, 화자의 개인성이 반영된 프레임들의 관측확률에 가중치를 주어 인식점수를 구하고 이를 이용하여 화자확인을 수행한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 문장종속형 화자확인 시스템에 대하여 살펴본다. 3장에서는 화자의 개인성 정보와 기존 인식방법의 문제점을 지적한 후, 본 논문에서 제안한 관측확률 가중방법과 이를 이용한 시스템의 구성에 대해서 설명하며, 4장에서 제안한 방법의 유효성을 검증하기 위한 실험방법과 결과를 보인 후, 5장에서 결론을 맺는다.

2 문장종속형 화자확인

화자확인의 목적은 주어진 음성으로 발성화자가 등록화자인지의 여부를 판단하는 것이다. 일반적인 화자확인 시스템의 구성은 그림 1과 같다. 먼저, 입력음성에서 음성신호에 포함되어 있는 개인성 정보를 나타내는 특징 파라미터를 벡터열의 형태로 추출한다. 특징 파라미터는 켈스트럼이나 인간의 청각적인 특성을 잘 반영한다고 알려져 있는 멜켈스트럼을 주로 사용한다. 학습 과정에서는 추출된 특징 벡터열을 이용하여 각각의 화자모델을 학습시킨다. 모델 학습 방법으로는 DTW(dynamic time warping), 신경 회로망(neural network), 벡터양자화(vector quantization), HMM등을 사용할 수 있다[2]. 화자 모델을 생성한 후, 적절한 임계치(threshold)를 정하면 학습과정은 끝나게 된다.

확인과정에서는 입력벡터열과 학습된 모델과의 유사도를 측정하여 인식점수를 구하고, 이를 앞의 학습과정에서 결정한 임계치와 비교하여 승인/거부 여부를 결정하게 된다. 학습방법으로는 주로 HMM을 사용하며, 이러한 시스템의 경우 일반적으로 유사도를 측정하는 방법으로 우도비 검사법을 사용한다.

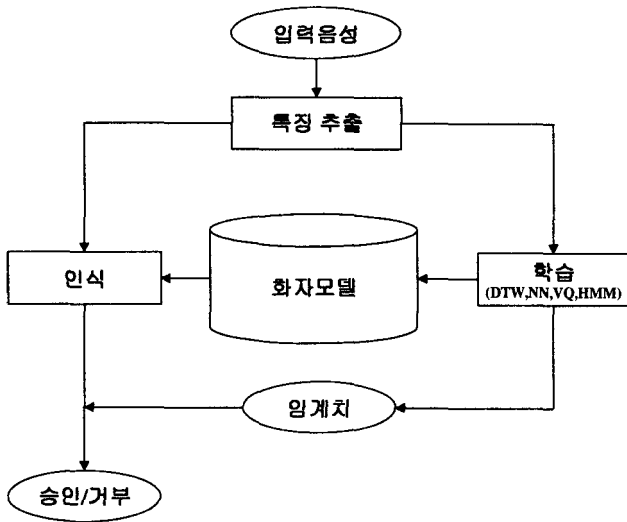


그림 1: 화자확인 시스템의 블럭도

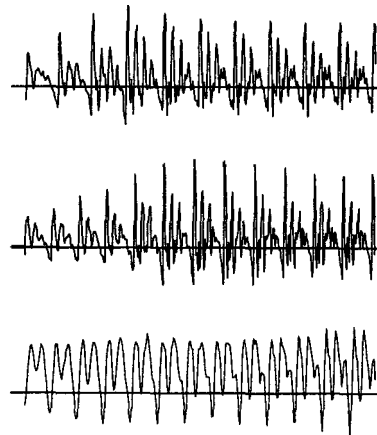


그림 2: 동일발성에 대한 화자간 변이

2.1 HMM기반 화자확인

HMM을 기반으로 하는 화자확인 시스템에서는 입력으로 주어지는 연속적인 음성을 프레임들로 나누고 각각에 대해 특징 벡터를 구한다. 따라서 입력음성은 특징 벡터의 이산적인 열로써 나타내어진다. \mathbf{X} 를 특징벡터의 열이라고 하고, T 를 음성신호의 전체 프레임의 수라고 하면, $\mathbf{X} = \{\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_T\}$ 로 나타낼 수 있다. HMM에서 i 번째 모델 λ_i 와 입력벡터열 \mathbf{X} 에 대하여, Viterbi 알고리즘을 이용하여 얻은 최적 상태열을 $S^i = \{s_1; s_2; \dots; s_T\}$ 라 하자. 이 경우 j 번째 상태에서의 관측 벡터 \mathbf{x} 에 대한 관측 확률은 $b_j^i(\mathbf{x}) = Pr(\mathbf{x}|s_t = j)$ 가 된다. 또 상태 j 에서 상태 k 로의 천이 확률 $a_{jk}^i = Pr(s_{t+1} = k|s_t = j)$ 로 정의된다. 상태 j 가 초기 상태가 될 확률을 $\pi_j = a_{0j}$ 로 정의하면 하나의 HMM 모델은 $\lambda_i = \{A, B, \pi\}$ 로 표현되어진다. 학습과정은 각 화자에 대한 모델 파라미터를 구하는 과정이 되어 전체 HMM 모델의 집합 $\Lambda = \{\lambda_i | 1 \leq i \leq \text{모델의 수}\}$ 를 얻는다[5].

화자확인 단계에서는 음성신호와 확인을 요구하는 화자모델을 이용하여, 둘 사이의 유사도를 측정하는 우도비 검사를 행하게 된다. 우도비 $f(\mathbf{X}, S^i | \lambda_i)$ 는 모델 i 에 대해 최적상태열과 특징 벡터열을 이용하여 유사도를 측정하는 것으로 시스템의 효율성을 위해 log를 취한 대수우도비(LLR; log-likelihood ratio)을 사용하고 길이 정규화를 위하여 프레임의 수로 나누어 준다. 이렇게 구한 인식 점수를 임계치와 비교하여 승인 및 거부 여부를 결정한다. 정규화된 대수우도비값은 학습과정에서 구한 HMM 모델 파라미터들을 이용하여 식 1과 같이 나타낼 수 있다[6].

$$g_i(\mathbf{X}; \Lambda) = \frac{1}{T} \log f(\mathbf{X}, S^i | \lambda_i) = \frac{1}{T} \left\{ \sum_{t=1}^T [\log a_{s_{t-1} s_t}^{(i)} + \log b_{s_t}^{(i)}(\mathbf{x}_t)] \right\} \quad (1)$$

$g_i(\mathbf{X}; \Lambda)$ 은 모델 i 에 벡터열 \mathbf{X} 를 입력하여 최적 상태열 S^i 에 대한 관측확률과 천이확률을 곱한 값들을 log를 취해서 프레임의 숫자로 정규화된 대수우도비값이다.

2.2 화자확인 시스템의 성능 평가

화자확인 시스템의 오류는 두가지로 나타내며 첫번째는 의뢰인을 거부하는 오인거부율(false rejection, type I error), 두번째는 사칭자를 수락하는 오인수락율(false acceptance, type II error)이 있다. 화자 확인에서 수락 및 거부의 결정은 식 2와 같이 확인 점수(verification score)에 특정 형태의 임계치를 적용함으로써 행해진다.

$$Result(\mathbf{X}, \lambda_i) = \begin{cases} \text{승인,} & \text{If } g_i(\mathbf{X}; \Lambda) > \text{임계치} \\ \text{거부,} & \text{otherwise} \end{cases} \quad (2)$$

임계치가 커질수록 오인거부율은 감소하지만 오인수락율은 증가하며, 임계치가 작아지면 그 반대로 증가, 감소한다. 따라서 화자확인 시스템의 성능은 일반적으로 오인거부율과 오인수락율이 같아지는 동일오류율(EER; equal error rate)로써 평가한다[3].

3 관측확률 가중기법

3.1 개인성이 포함된 프레임의 간단한 소개 및 선택방법

화자의 개인성 정보는 발생 기관의 해부학적 구조 차이에 기인하는 선천적인 특성과 개인의 발생 습관으로 대표되는 후천적인 특성으로 나누어 볼 수 있다. 이중 화자의 동적특성을 반영하는 후천적 특징이란 화자의 억양, 강세, 빠르기 등과 같은 발생 습관을 말하며 음향 파라미터의 변화 형태로부터 관측될 수 있다. 그림 2의 (a), (b)는 동일화자가 발생한 음성의 파형이며, (c)는 다른 화자가 발생한 동일 음성의 파형이다. 그림에서 보는 바와 같이 동일한 문장에 대해서도 화자간의 개

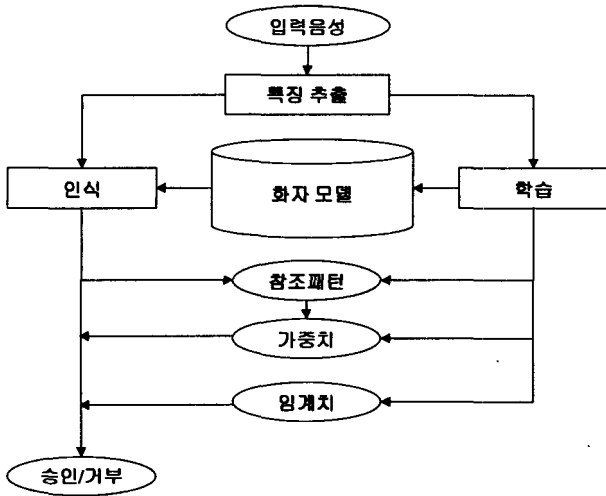


그림 3: 제안한 화자확인 시스템의 블록도

인차에 따라 크게 다른 형태로 나타나게 된다. 음성인식의 경우, 이러한 변이를 흡수하여 서로 다른 화자에 대해서도 인식이 될 수 있도록 적응시켜야 하지만, 화자인식에서는 이러한 차이를 이용하여 화자들을 구분해야 한다. 하지만 기존에 음성인식에서의 방법을 그대로 사용하는 경우, 이러한 개인차를 효과적으로 반영할 수 없다.

본 논문에서는 다음과 같은 방법을 제안하여 이러한 문제점을 해결한다. 학습과정에서 화자모델 i 의 학습자료들에 대하여 특징을 추출한 후, 각 프레임 비교한다. 이때 프레임간의 거리를 적절한 거리척도를 이용하여 먼 것과 가까운 것을 구분한다. 학습자료들간에 정합되는 프레임들 사이의 거리가 멀다는 것은 화자내에서의 변이가 있을 수 있다는 의미로 해석될 수 있다. 반면에 프레임간 거리가 가까운 프레임들은 화자내에서 변이가 적으므로 이 부분에 화자의 개인성 정보가 포함되어 있다고 가정할 수 있다. 프레임간의 거리를 이용하여 각 프레임에 대한 가중 척도를 계산하는 방법은 3.3절에서 설명한다.

3.2 시스템의 구성

제안된 시스템의 구성은 그림 3과 같이 구성된다. 학습과정에서 기존의 시스템에 각각의 모델에 대하여 프레임별 가중치를 구하는 과정이 새롭게 추가된다. 가중치를 구하는 방법은 주어진 N 개의 학습자료들 중 인식점수가 가장 좋은 학습자료 하나를 참조패턴으로 선택한 후, 나머지 학습자료들을 이 참조패턴과 DTW 탐색을 이용하여 프레임들을 정합시킨다. 참조패턴에 대하여 각각의 학습자료들과 정합되는 프레임에 대한 거리를 구한 후, 이 값을 이용하여 가중치를 계산한다.

인식과정에서는 화자가 발성한 음성에 대하여 특징파라미터를 추출한 후, Viterbi 탐색에 의하여 HMM 모델에 대한 최적의 경로를 찾아 최적상태열을 얻는다. 학습과정에서 얻은 참조패턴과 입력패턴을 DTW 알고리즘을 이용하여 최적의 정합열을 찾고, 각각의 프레임에 대한 가중치를 얻는다. 최적상태열과 가중치를 얻은 다음, 각각의 상태의 관측확률에 그 상

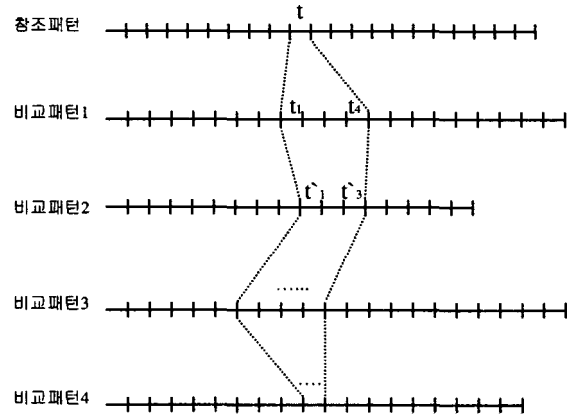


그림 4: 참조패턴과 비교패턴의 정합

태에 해당하는 프레임의 가중치를 곱하여 새로운 출현확률을 구한다. 이 과정을 통하여 얻은 인식 점수를 임계치와 비교하여 승인 및 거부 여부를 결정한다.

3.3 관측확률 가중

일반적인 HMM을 기반으로 하는 화자확인 시스템에서의 대수우도비값은 2.1절의 식 1과 같다. 시스템에 인식을 위한 음성이 입력되면 먼저, 특징 파라미터를 추출한 후, 저장되어 있는 참조 패턴과 비교하여 최적의 정합열을 구한다. 그리고, 이를 이용하여 각 프레임에 대한 가중치를 계산한다. Viterbi 탐색에 의해 주어진 입력음성에 대한 최적의 상태열을 얻은 후, 앞에서 계산한 프레임별 가중치 값을 각각의 프레임과 정합되는 상태의 관측확률에 곱해서 새로운 우도비점수를 식 3과 같이 구할 수 있다.

$$\begin{aligned}
 g'_i(\mathbf{X}; A) &= \frac{1}{T} \log f'(\mathbf{X}, S^i | \lambda_i) \\
 &= \frac{1}{T} \left\{ \sum_{t=1}^T [\log a_{s_{t-1}s_t}^{(i)} + \log (b_{s_t}^{(i)}(\mathbf{x}_t) \times w^{(i)}(t))] \right\} \quad (3) \\
 &= \frac{1}{T} \left\{ \sum_{t=1}^T [\log a_{s_{t-1}s_t}^{(i)} + \log b_{s_t}^{(i)}(\mathbf{x}_t) + \log w^{(i)}(t)] \right\}
 \end{aligned}$$

여기서 $w^{(i)}(t)$ 는 모델 i 에서 프레임 t 에 대한 가중치를 말하며, 식 4와 같이 구할 수 있다.

$$w^{(i)}(t) = \frac{1}{\frac{1}{N-1} \sum_{n \neq ref} \delta(\mathbf{x}_t^{ref}, \mathbf{x}_{t_1, \dots, t_M}^n)} \quad (4)$$

N 은 학습에 사용되는 학습자료의 수이며, $\delta(\mathbf{x}_t^{ref}, \mathbf{x}_{t_1, \dots, t_M}^n)$ 은 참조 패턴의 t 번째 프레임과 이에 정합되는 n 번째 학습자료의 t_1 프레임에서 t_M 프레임까지의 거리를 나타낸다. \mathbf{x}^{ref} 와 \mathbf{x}^n 의 DTW 탐색을 수행한 후에 정합되는 프레임들을 보면, 그림 4에서와 같이 \mathbf{x}_t^{ref} 에 정합되는 \mathbf{x}^n 의 프레임이 여러개일 수 있으며, 반대로 \mathbf{x}^{ref} 의 여러 프레임이 \mathbf{x}^n 의 한 프레임에 정합될 수도 있다. $\delta(\cdot)$ 의 값은 참조패턴에

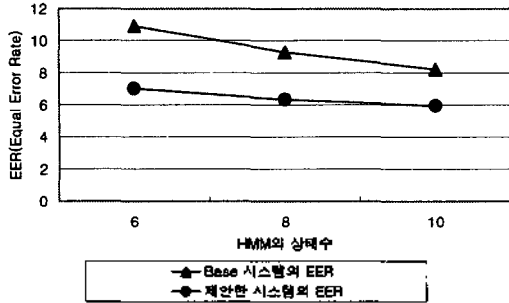


그림 5: HMM 상태수의 따른 동일오류율(EER)

맞추어지므로, 정합되는 입력패턴의 수가 많은 경우, 평균값을 x_i^{ref} 의 값으로 하고, 입력패턴 하나가 참조패턴 여러개와 정합되는 경우에는 각각의 거리를 계산하게 된다.

$$d(x_i^{ref}, x_{i_1, \dots, i_M}^n) = \frac{1}{M} \sum_{i=1}^M |x_i^{ref} - x_{i_i}^n| \quad (5)$$

가중치를 각각의 프레임에 할당하는 경우, over-fitting의 우려가 있어 10프레임을 하나의 단위로 묶어서 거리를 구하고, 가중치를 정하였다. 거리척도로는 프레임간 거리의 평균값이나 표준편차 등의 Euclidean 거리를 사용하여 구할 수 있다.

4 결과

4.1 실험환경

제안한 방법의 유효성을 검증하기 위해 고립단어에 대해 화자확인 성능을 비교하였다. 실험에서 사용된 자료는 12명의 화자들이 5가지 종류의 4자리 숫자를 5번씩 발성한 총 300개의 자료를 사용하여 60개의 화자모델을 얻었다. 또한, 동일한 화자들이 4번에 나누어서 학습과정과 같은 숫자음을 한번에 2번씩, 8번 발성한 총 480개의 발성 자료를 실험에 사용하였다. 특징 파라미터는 인간의 청각특성을 반영한 13차 멜단위 캡스트럼 벡터를 사용하였다. 인식시스템은 상태마다 코드북을 가지고 그 왜곡거리를 출력확률로 계산하는 HMVQM으로 구현하였으며, 선형연결구조(left-to-right topology)로 구성하였다[1]. 또한 상태수에 따른 성능의 변이를 알아보기 위해 6, 8, 10 세가지로 나누어 비교실험하였다. 기본 시스템은 기존의 우도비 점수를 사용하는 HMVQM 시스템으로 구성하였으며, 제안한 시스템은 여기에 관측확률 가중기법을 적용하여 구현하였다. 거리 척도는 프레임간 거리의 Euclidean 거리를 사용하였다.

4.2 실험결과

표 1은 HMM의 상태수를 6, 8, 10으로 변화시킴에 따른 기존시스템과 제안한 시스템의 오인수락율(FA)과 오인거부율(FR)을 보인 것이다. 그림 5는 기존 시스템과 제안한 시스템의 동일오류율(EER)을 보인 것이다. 그림 5에서 보는 바와

표 1: HMM 상태수별 오인수락율(FA)와 오인거부율(FR)

실험	상태수 6	상태수 8	상태수 10
base 시스템(FA)	10.59%	9.22%	8.08%
제안한 시스템(FA)	7.15%	6.19%	5.85%
base 시스템(FR)	11.25%	9.38%	8.33%
제안한 시스템(FR)	6.88%	6.46%	6.04%

같이 HMM의 상태수에 관계없이 제안한 시스템이 기존의 시스템보다 동일오류율(EER)이 2~3%정도 감소한 것을 알 수 있다. 동일오류율(EER)의 감소로 화자확인 시스템의 전체적인 인식율이 향상되었다. 제안한 시스템의 HMM 상태수가 늘어남에 따라 경우 성능향상의 폭이 기존 시스템에 비해 크지 않음을 통해 화자의 개인특성이 반영되었음을 알 수 있다.

5 결론

본 연구에서는 HMM기반 문장중속형 화자확인 시스템의 성능 개선을 위하여 음성에 포함되어 있는 화자의 개인성을 반영한 인식점수를 계산하는 방법을 제안하였다. 기존의 시스템에 비해 동일오류율(EER)이 평균 3% 감소하였으며, 이로 인해 전체 인식율이 향상되어 제안한 방법이 화자확인에 더 적합함을 보였다. 따라서 화자인식시스템의 성능향상을 위해서는 화자의 개인성 정보의 효율적인 추출과 사용이 중요하다는 것을 알 수 있었다. 현재는 화자의 개인성을 화자모델에 직접 적용할 수 있는 학습법과 임계치결정에 대한 연구를 진행 중이다.

참고 문헌

- [1] 윤성진. 적은 학습자료 환경하에서 화자인식 시스템의 성능향상에 관한 연구. 석사학위 논문, 한국과학기술원, 1994.
- [2] 이광수. 화자인식기술. 제15회 음성통신 및 신호처리 워크샵논문집, pages 42-46, 1995.
- [3] Sadaoki Furui and M. Mohan Sondhi. Advances in Speech Signal Processing. Marcel Dekker Inc., 1992.
- [4] Douglas O'Shaughnessy. Speaker Recognition. *IEEE ASSP Magazine*, pages 4-17, 10 1986.
- [5] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257-286, 2 1989.
- [6] Aaron E. Rogenberg and Chin-Hui Lee. The Use of Cohort Normalized Scores for Speaker Recognition. In *Proceedings of ICSLP*, pages 599-602, 1992.