개인성 정보의 가중화에 의한 화자확인의 성능향상

김 세 현○ 장 길 진 오 영 환

한국과학기술원 전산학과

shkim@bulsai.kaist.ac.kr

Performance Improvement of Speaker Verification System By Speaker Information Weighting

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

Department of Computer Science

Korea Advanced Institute of Science and Technology

요 약

기존의 문장종속형 화자인식 기법에서는 음성 신호의 각 분석 프레임이 같은 기여도를 갖는 것으로 간주한다. 화자인식 시스템의 성능향상을 위해서는 음운정보보다는 인식의 단서가 되는 화자의 개인성 정보가 잘 반영되도록 하는 것이 중요하다. 본 논문에서는 HMM (hidden Markov model)을 기반으로 한 문장종속형 화자확인 시스템의 성능향상을 위해 프레임별로 인식의 단서가 되는 개인성 정보의 양을 측정하는 방법과, 이를 화자확인 시스템에 적용하는 기법을 제안한다. 제안한 방법을 적용한 결과, 기존의 우도비(likelihood ratio) 정규화 점수를 사용하는 방법에 비해 동일오류율(EER; equal error rate)을 평균 34% 감소시켜 인식율 향상을 얻을 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]. 화자확인은 발성된 음성이 원하는 화자인지 아닌지를 구분해 내는 것이며, 화자식별은 등록된 목적 화자들 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 화자확인에서는 음성 신호의 각 프레임이 같은 기여도를 갖는 것으로 보고 인식을 수행한다. 그러나, 음성신호의 각 구간이 인식의 단서를 제공하는 정도가 다르므로 이를 측정하여 사용하는 인식 방법들이 필요하게 된다. 화자인식의 경우, 인식의 단서는 화자의 개인성 정보이며, 이 정보를 이용하여 화자의 특징을 나타낸다. 제안하는 방법에서는 주어진 학습자료를 이용하여 음성의 프레임에 포함되어 있는 개인성 정보의 양을 측정한 후, 측정값을 프레임들의 관측확률에 가중치로 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 개인성이 포함된 프레임을 선택하는 방법과 이를 이용한 관측확률에 대한 가중방법 및 시스템의 구성에 대해서 설명하며, 4장에서 제안한 방법의 유효성을 검증하기 위한 실험방법과 결과를 보인 후, 5장에서 결론을 맺는다.

2. 문장종속형 화자확인

일반적인 화자확인의 과정은 다음과 같다. 먼저, 입력음성에서 음성신호에 포함되어 있는 개인성 정보를 나타내는 특징 파라미터를 벡터열의 형태로 추출한다. 학습 과정에서는 추출된 특징 벡터열을 이용하여 각각의 화자모델을 학습시킨다. 모델 학습 방법으로는 DTW (dynamic time warping), 신경회로망(neural network), 벡터양자화(vector quantization), HMM 등을 사용할 수 있다. 화자 모델을 생성한 후, 적절한 임계치(threshold)를 정하면 학습과정은 종료된다.

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

2.1 HMM기반 화자확인

HMM을 기반으로 하는 화자확인 시스템에서는 입력으로 주어지는 연속적인 음성을 프레임들로 나누고 각각에 대해 특징 벡터를 구한다. 입력음성은 특징 벡터의 이산적인 열로써 나타내어진다. $\mathbf{X}$를 특징벡터의 열이라고 하고, T를 음성신호의 전체 프레임 수라고 하면, $\mathbf{X} = \{\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_T\}$로 나타낼 수 있다. HMM에서 i번째 모델 λ_i와 입력벡터열 $\mathbf{X}$에 대하여, Viterbi 알고리즘을 이용하여 얻은 최적 상태열을 $S^i = \{s_1; s_2; \cdots; s_T\}$라 하자. 이 경우 j번째 상태에서의 관측 벡터 $\mathbf{x}$에 대한 관측확률은 $b_j^i(\mathbf{x}) = Pr(\mathbf{x}|s_t = j)$가 된다. 또 상태 j에서 상태 k로의 천이 확률 $a_{jk}^i = Pr(s_{t+1} = k|s_t = j)$로 정의된다. 상태 j가 초기 상태가 될 확률을 $\pi_j = a_{0j}$로 정의하면 하나의 HMM 모

델은 $\lambda_i = \{A, B, \pi\}$로 표현되어진다. 학습과정은 각 화자에 대한 모델 파라미터를 구하는 과정이 되어 전체 HMM 모델의 집합 $\Lambda = \{\lambda_i | 1 \leq i \leq$ 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 i에 대해 최적상태열과 특징 벡터열을 이용하여 유사도를 측정하는 것으로 시스템의 효율성을 위해 log를 취한 대수우도비(LLR; log-likelihood ratio)을 사용하고 길이 정규화를 위하여 이 값을 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.

$$\begin{aligned} g_i(\mathbf{X};\Lambda) &= \frac{1}{T}\log f(\mathbf{X}, S^i|\lambda_i) \\ &= \frac{1}{T}\{\sum_{t=1}^{T}[\log a^{(i)}_{s_{t-1}s_t} + \log b^{(i)}_{s_t}(\mathbf{x}_t)]\} \end{aligned} \quad (1)$$

$g_i(\mathbf{X};\Lambda)$은 모델 i에 벡터열 $\mathbf{X}$를 입력하여 최적 상태열 S^i에 대한 관측확률과 천이확률을 곱한 값들을 log를 취해서 프레임 수로 정규화한 대수우도비값이다.

2.2 화자확인 시스템의 성능 평가

화자확인 시스템에는 두가지 종류의 오류가 발생한다. 첫번째는 의뢰인을 거부하는 오인거부율(false rejection, type I error)이며, 두번째는 사칭자를 수락하는 오인수락율(false acceptance, type II error)이다. 화자 확인에서 수락 및 거부의 결정은 확인 점수(verification score)에 특정 형태의 임계치를 적용함으로써 행해진다. 일반적으로 임계치가 커질수록 오인거부율은 감소하지만 오인수락율은 증가하며, 임계치가 작아지면 그 반대로 각각 증가, 감소한다. 따라서 화자확인 시스템의 성능은 일반적으로 오인거부율과 오인수락율이 같아지는 동일오류율(EER; equal error rate)로써 평가한다 [3].

3. 개인성 정보량의 측정과 가중기법

앞장에 살펴본 바와 같이 통상적인 화자확인 시스템에서는 음성의 모든 구간에 같은 비중을 두고 인식 점수를 구한다. 그러나, 음성 각 구간이 인식의 단서를 제공하는 정도가 다르므로, 이러한 음성 구간들의 차이를 측정하고, 이를 이용하는 방법이 필요하다. 화자인식에서 인식의 단서는 화자의 개인성 정보이다. 따라서 본 논문에서는 음성구간에 포함되어 있는 화자의 개인성 정보의 양을 측정하고, 이를 인식점수에 반영하는 방법을 제안한다.

3.1 개인성 정보와 정보량의 측정

화자의 개인성 정보는 발성 기관의 해부학적 구조 차이에 기인하는 선천적인 특성과 개인의 발성 습관으로 대표되는 후천적인 특성으로 나누어 볼 수 있다. 이중 화자의 동적특성을 반영하는 후천적 특징이란 화자의 억양, 강세, 빠르기 등과 같은 발성 습관을 말하며 음향 파라미터의 변화 형태로부터 관측될 수 있다. 화자의 개인차에 의해서 동일한 발성에 대해서도 화자간에 변이가 발생하게 되는데, 음성인식의 경우에는 이러한 변이를 흡수하여 서로 다른 화자에 대해서도 인식이 될 수 있도록 적응시켜야 하지만, 화자인식에서는 이러한 차이를 이용하여 화자들을 구분해야 한다. 일반적으로 특징파라미터는 다음의 식으로 나타나는 F-비를 이용하여 선정한다 [2]. 이 값이 클수록 화자의 개인성 정보를 보다 잘 반영하고 있는 특징파라미터로 간주한다.

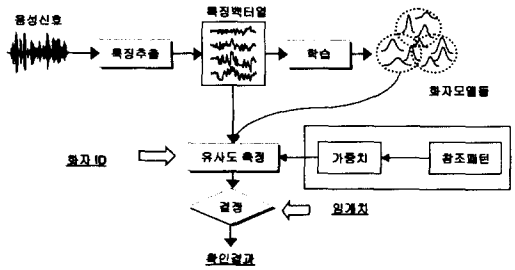

그림 1: 제안한 화자확인 시스템의 블럭도

$$\text{F-ratio} = \frac{\text{inter-speaker variance}}{\text{intra-speaker variance}} \quad (2)$$

즉, 화자간의 변이는 크게 하고 화자내의 변이는 작게 하는 부분이 화자의 개인성이 잘 반영되는 부분이다.

본 논문에서는 화자가 발성한 음성의 프레임별로 F-비를 측정하여 각 프레임마다 화자의 개인성 정보의 반영정도를 정의한다. 그리고 이 값을 가중치로 이용하여 인식을 수행하는 방법을 제안한다. 화자의 개인성 정보의 양을 결정하는 방법은 다음과 같다. 학습과정에서 화자모델 i의 학습자료들과 반화자모델들에 대하여 특징을 추출한 후, 각 프레임을 비교한다. 이때 프레임간의 거리를 적절한 거리척도를 이용하여 먼 것과 가까운 것으로 구분한다. 화자모델들의 학습자료들간의 거리를 측정하여 이를 화자내의 변이값으로 하고, 반화자모델과의 거리를 측정하여 이를 화자간의 변이값으로 한다. 이 두 값을 이용하여 프레임별로 F-비를 구해보면, 값이 큰 프레임은 개인성 정보가 많이 포함되어 있음을 의미하며, 값이 작은 것은 상대적으로 적은 양의 정보가 포함되어 있다는 것을 뜻한다. 이 값을 이용하여 가중 척도를 계산하는 방법은 다음 절에서 설명한다.

3.2 화자확인에서의 가중방법

앞장에서 살펴본 바와 같이, HMM을 기반으로 하는 화자확인 시스템에서는 인식점수 계산을 위해서 천이확률과 프레임의 관측확률을 이용한다. 본 논문에서 제안된 시스템의 구성은 그림 1과 같다. 학습과정에 기존의 시스템에 각각의 모델에 대하여 프레임별 가중치를 구하는 과정이 새롭게 추가된다. 가중치를 구하는 방법은 주어진 N개의 학습자료들 중 인식점수가 가장 좋은 학습자료 하나를 참조패턴으로 선택한 후, 나머지 학습자료들과 반화자모델들의 학습자료를 이 참조패턴과 DTW 탐색을 이용하여 프레임들을 정합시킨다. 참조패턴에 대하여 각각의 학습자료들과 정합되는 프레임에 대한 거리를 구한 후, 이 값을 이용하여 프레임별 F-비를 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 상태에 해당하는 프레임의 가중치를 가중하여 새로운 관측확률을 구한다. 이 과정을 통하여 얻은 인식 점수를 임계치와 비교하여 승인 및 거부 여부를 결정한다.

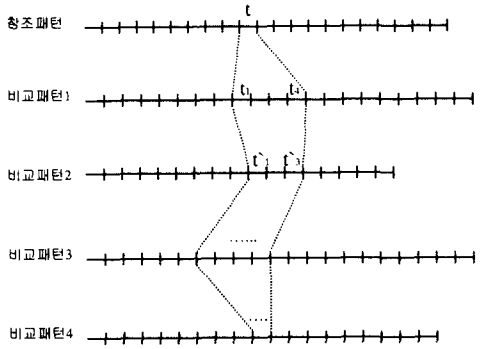

그림 2: 참조패턴과 비교패턴의 정합

일반적인 HMM을 기반으로 하는 화자확인 시스템에서의 대수 우도비값은 2.1절의 식 1과 같다. 여기에 화자의 개인성 정보를 반영하기 위하여, 3.1절에서 제안한 방법대로 측정한 프레임별 화자의 개인성 정보량을 가중한 시스템의 인식 점수는 식 3과 같이 구할 수 있다.

$$\begin{aligned} g_i^{new}(\mathbf{X};\Lambda) &= \frac{1}{T}\log f^{new}(\mathbf{X}, S^i|\lambda_i) \\ &= \frac{1}{T}\{\sum_{t=1}^{T}[\log a^{(i)}_{s_{t-1}s_t} + \log\{b^{(i)}_{s_t}(\mathbf{x}_t)^{-w_t^{(i)}}\}]\} \\ &= \frac{1}{T}\{\sum_{t=1}^{T}[\log a^{(i)}_{s_{t-1}s_t} - w_t^{(i)}\log b^{(i)}_{s_t}(\mathbf{x}_t)]\} \end{aligned} \quad (3)$$

여기서 $w^{(i)}(t)$는 모델 i에서 프레임 t에 대한 가중치를 말하며, 가중을 위한 프레임의 출력확률값이 0과 1사이의 값이므로 -1을 곱해 가중하게 된다. $w^{(i)}(t)$를 구하는 식은 다음과 같다.

$$w_t^{(i)} = \frac{\frac{1}{M}\sum_{all\ m}\delta(\mathbf{x}_t^{ref},\ \mathbf{x}^m_{t_1;\cdots;t_J})}{\frac{1}{N-1}\sum_{n\neq ref}\delta(\mathbf{x}_t^{ref},\ \mathbf{x}^n_{t_1;\cdots;t_K})} \quad (4)$$

$$\delta(\mathbf{x}_t^{ref},\ \mathbf{x}^n_{t_1;\cdots;t_K}) = \frac{1}{K}\sum_{i=1}^{K}\|\mathbf{x}_t^{ref} - \mathbf{x}^n_{t_i}\|_\epsilon \quad (5)$$

분모항은 화자모델의 학습자료와의 거리를 나타내며, 분자항은 반화자모델 자료와의 거리를 나타낸다. $\delta(\mathbf{x}_t^{ref},\ \mathbf{x}^n_{t_1;\cdots;t_K})$은 참조패턴의 t번째 프레임과 이에 정합되는 n번째 자료의 t_1 프레임에서 t_K 프레임 사이의 거리를 나타낸다. $\mathbf{x}^{ref}$와 $\mathbf{x}^n$의 DTW 탐색을 수행한 후에 정합되는 프레임들을 보면, 그림 2에서와 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 맞추어지므로, 정합되는 입력패턴의 수가 많은 경우, 평균값을 $\mathbf{x}_t^{ref}$의 값으로 하고, 입력패턴 하나가 참조패턴 여러개와 정합되는 경우에는 각각의 거리를 계산하게 된다.

4. 결 과

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 멜단위 켑스트럼 벡터를 사용하였다. 인식시스템은 상태마다 코드북을 가지고 그 왜곡거리를 출력확률로 계산하는 HMVQM으로 구현하였다 [1]. 또한 상태수에 따른 성능의 변이를 알아보기 위해 6, 8, 10 세가지로 나누어 비교실험하였다. 기본 시스템은 기존의 우도비 점수를 사용하는 HMVQM 시스템으로 구성하였다.

표 1: HMM 상태수별 동일오류율(EER)

실험		상태수 6	상태수 8	상태수 10
base 시스템	FA	10.59%	9.22%	8.08%
	FR	11.25%	9.38%	8.33%
	EER	10.92%	9.3%	8.205%
제안한 시스템	FA	6.76%	5.82%	5.85%
	FR	7.08%	6.46%	5.42%
	EER	6.92%	6.14%	5.635%

표 1은 HMM의 상태수를 6, 8, 10으로 변화시킴에 따른 기존시스템과 제안한 시스템의 오인수락율(FA)과 오인거부율(FR)과 동일오류율(EER)을 보인것이다. 표에서 보는 바와 같이 HMM의 상태수에 관계없이 제안한 시스템의 동일오류율이 기존의 시스템보다 33~34%정도 감소한 것을 알 수있다. 동일오류율의 감소로 화자확인 시스템의 전체적인 인식율이 향상되었다. 제안한 시스템의 HMM 상태수와 관계없이 성능이 향상되었음을 통해 프레임별로 화자의 개인성 정보 반영 정도를 달리하는 인식 방법이 효과적임을 알 수 있었다.

5. 결론

본 연구에서는 HMM기반 문장종속형 화자확인 시스템의 성능개선을 위하여 음성에 포함되어 있는 화자의 개인성 정보의 양을 측정하는 방법과 이를 반영하여 인식점수를 계산하는 방법을 제안하였다. 기존의 시스템에 비해 동일오류율(EER)이 평균 34% 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 반화자모델에 사용할 배경화자를 선택하기 위한 방법에 대한 연구를 진행중이다.

참고 문헌

[1] 윤성진, 적은 학습자료 환경하에서 화자인식 시스템의 성능향상에 관한 연구, 석사학위 논문, 한국과학기술원, 1994.

[2] S. Furui, *Digital speech processing, synthesis, and recognition*, Marcel Dekker Inc, 1992.

[3] S. Furui and M. M. Sondhi, *Advances in Speech Signal Processing*, Marcel Dekker Inc., 1992.

[4] D. O'Shaughnessy, *Speaker Recognition*, IEEE ASSP Magazine, (1986), pp. 4-17.

[5] L. R. Rabiner, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Processings of the IEEE, 77 (1989), pp. 257-286.