말소리와 음성과학 제1권 제1호 (2009)
pp. 55~62

핵심어 인식기에서 단어의 음소레벨 로그 우도 비율의 패턴을 이용한 발화검증 방법

Utterance Verification using Phone-Level Log-Likelihood Ratio Patterns in Word Spotting Systems

김 정 현[1] · 권 석 봉[2] · 김 회 린[3]
Kim, Chong-Hyon · Kwon, Suk-Bong · Kim, Hoi-Rin

ABSTRACT

This paper proposes an improved method to verify a keyword segment that results from a word spotting system. First a baseline word spotting system is implemented. In order to improve performance of the word spotting systems, we use a two-pass structure which consists of a word spotting system and an utterance verification system. Using the basic likelihood ratio test (LRT) based utterance verification system to verify the keywords, there have been certain problems which lead to performance degradation. So, we propose a method which uses phone-level log-likelihood ratios (PLLR) patterns in computing confidence measures for each keyword. The proposed method generates weights according to the PLLR patterns and assigns different weights to each phone in the process of generating confidence measures for the keywords. This proposed method has shown to be more appropriate to word spotting systems and we can achieve improvement in final word spotting accuracy.

Keywords: Utterance verification, word spotting, PLLR pattern

1. 서 론

현재 여러 응용분야에서 음성인식의 중요성은 증가하고 있는 추세이다. 최근에 홈 오토메이션, 자동차 네비게이션 시스템 등에 대한 관심의 증가로 이와 같은 시스템에서 믿을 만한 성능을 보이며 동작하는 음성인식 시스템이 필요하게 되었다. 특히 인간의 음성은 연속적이라는 특성을 지니기 때문에 연속 입력음성에 대한 인식은 더욱 중요하다고 할 수 있다. 이와 같이 연속입력음성에서 사용자에 의해 정의된 특정단어 혹은 어구 등을 검색하는 시스템을 핵심어 인식 시스템이라고 한다. 일반적인 단어 인식 시스템이 입력음성 전체에 대한 인식결과를 출력하는데 반해 핵심어 인식 시스템은 입력음성에서 사용자에게 의미있는 특정 부분에 대한 인식 결과를 출력하는 것을 목적으로 한다. 이와 같은 핵심어 인식 시스템에 대해 과거로부터 많은 연구가 이루어져 왔으며 특히 dynamic time warping (DTW)에 기반을 둔 template matching 방식 [1], 현재 널리 사용되고 있는 hidden Markov model (HMM) 을 기반으로 한 방식 [2] 등이 주로 사용되어 왔다. 본 논문에서는 HMM 을 기반으로 한 핵심어 인식 시스템을 구현하였다. HMM 을 기반으로 한 핵심어 인식 시스템은 maximum likelihood (ML) 훈련을 통해 다양한 화자 및 단어에 따른 특성을 흡수한 HMM 을 이용하여 template 기반 방식에 비해 핵심어 음성뿐만 아니라 비 핵심어 음성에 대한 모델링을 더욱 적절히 할 수 있다는 특성을 지닌다 [2]. 이와 같이 HMM 을 이용하여 핵심어 및 비 핵심어에 대한 모델링을 하더라고 핵심어 인식 시스템만으로는 좋은 성능을 얻기 힘들며 이와 같은 특성에 의해 핵심어 인식 시스템 후단에 발화검증 시스템을 추가시켜 사용하는 방식을 사용하고 있다 [3].

발화검증은 음성인식 결과의 신뢰도를 신뢰도 척도라는 값을 통해 결정하는 기술이다. 음성인식기의 입력음성이 항상 인식대상어휘에 해당하는 음성이라고 볼 수 없고 또한 발화검증 시스템을 제외한 음성인식시스템의 인식결과가 항상 올바르다

1) 한국과학기술원 polarbear@icu.ac.kr, 교신저자
2) 한국과학기술원 sbkwon@icu.ac.kr
3) 한국과학기술원 hrkim@ee.kaist.ac.kr
(이 논문은 연구지원재단의 지원금으로 수행된 연구입니다.(지원번호: KOR-1234-56798))

접수일자: 2009년 1월 31일
수정일자: 2009년 3월 10일
게재결정: 2009년 3월 15일

고 볼 수 없기 때문에 발화검증은 음성인식에서 중요한 부분을 차지하고 있다. 발화검증에서는 신뢰도 척도를 구하고 이를 미리 정한 임계치와 비교하여 인식결과의 수락 / 거절을 결정한다. 신뢰도 척도로는 LRT 기반의 신뢰도 척도가 가장 많이 사용되어 왔고 이후 이와 같은 신뢰도 척도를 기반으로 minimum classification error (MCE), minimum verification error (MVE) 등의 discriminative training을 통하여 반 가설을 모델링하여 성능 개선을 하는 방법들 [4]이 제안되었다. 현재에는 기존의 하나의 신뢰도척도만을 사용하는 방식에서 벗어나 N-best list 내의 후보 인식결과, bayesian 등을 사용하여 다양한 신뢰도 척도를 구한 후, 이를 통합하여 사용하고 있다.

음소단위의 인식을 하는 경우 기존의 LRT 기반의 기본적인 방식은 다음과 같은 방식으로 각 단어에 대한 신뢰도 척도를 구한다. 먼저 입력음성을 음소단위로 인식한다. 인식된 각 음소에 대해 가설검정을 수행하여 음소단위의 신뢰도 척도를 구한다. 이와 같이 구한 음소단위의 신뢰도 척도의 평균을 구하여 통합해 단어에 대한 신뢰도 척도를 구한다 [6]. 기본적인 LRT 방식을 핵심어 인식시스템에 사용하여 발화검증을 수행한 경우 성능저하로 이어지는 몇 가지 문제점이 발생하였다. 이러한 문제점을 해결하기 위해 본 논문에서는 단어의 PLLR 패턴을 이용한 발화검증 방식을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 논문에 사용된 기본적인 HMM 기반 핵심어 인식 시스템 및 LRT 기반의 발화검증 방식에 대해 설명한다. 또한 기존의 방식 중 하나인 Word Voiceprint 방식을 소개한다. 3장에서는 기존의 LRT 기반 발화검증방식을 핵심어 인식시스템에 사용한 경우 문제점에 대해 설명하고 제안된 방식인 단어의 PLLR 패턴을 이용한 발화검증 방식을 설명한다. 4장에서는 기존 방식 및 제안된 방식을 사용하여 발화검증 및 핵심어 인식을 수행한 실험결과를 보이고, 5장에서 결론을 맺는다.

2 HMM 기반 핵심어 인식 및 LRT 기반의 발화검증 시스템

2.1 HMM을 기반으로 한 핵심어인식 시스템

본 논문에서는 HMM을 기반으로 한 핵심어 인식 시스템을 구현하였다. 이와 같은 핵심어 인식 시스템은 핵심어에 대한 모델 그리고 비 핵심어에 대한 모델 즉 garbage 모델의 두 가지 모델로 구성된다. 본 논문에서는 sub-word 단위를 기반으로 각 모델을 모델링 하였다. Sub-word 단위로 모델링하는 경우 시스템은 핵심어 혹은 시스템 변경이 용이하다는 특성을 지닌다.

핵심어에 대한 모델의 기본 단위로는 문맥 종속형(context-dependent) 모델 중 triphone을 사용하였고 각 triphone들을 결합하여 핵심어를 모델링 한다. garbage 모델의 기본단위로는 문맥 독립형 (context-independent) 모델 중 monophone을 사용하였고 단일 monophone을 사용하여 각 garbage 모델을 모델링하여 핵심어 모델에 비해서는 덜 정교하지만 신뢰성 있게 모델링한다. <그림1>은 구현한 핵심어 인식 시스템을 나타낸다.

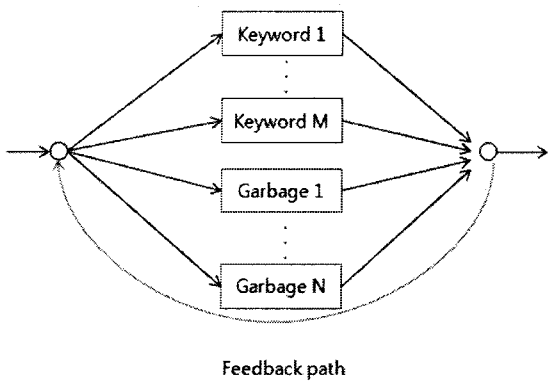

그림 1. 핵심어 인식 시스템의 구조
Figure 1. Structure of the word spotting system

이와 같은 시스템에서 각 핵심어 및 garbage 모델은 동일한 확률을 갖고 연결되며 전체적인 네트워크는 루프를 형성하여 핵심어 및 garbage 모델의 sequence로 입력음성을 모델링한다. 입력음성이 핵심어를 포함하는 경우, 핵심어 모델이 그에 해당하는 garbage 모델 sequence에 비해 큰 우도를 누적하게 되고 비터비 디코딩시 핵심어를 포함하는 경로를 디코딩할 가능성이 높게 되어 이에 따라 핵심어를 검출할 수 있게 된다.

핵심어 인식 시스템에서는 음성 내에 존재하는 핵심어를 인식하지 못한 경우와 핵심어에 해당하지 않은 음성의 부분을 핵심어로 인식한 경우의 두 가지 오류가 존재한다. 본 논문에서는 핵심어 모델이 garbage 모델에 비해 단어 간 천이가 덜 빈번하다는 점을 이용해 두 오류의 비율을 word insertion penalty 값을 사용해 조정한다. word insertion penalty는 비터비 탐색시 토큰이 하나의 단어에서 다른 단어로 이동할 때 부가하는 값을 의미한다. 핵심어 및 비 핵심어에 부분에 대한 모델링을 적절히 한다고 하여도 이와 같은 핵심어 인식 시스템만으로는 좋은 성능을 얻기 힘들며 후단에 발화검증 시스템을 추가하여 성능향상을 위한 방안으로 사용한다 [3]. <그림2>에 핵심어 인식에 이은 발화검증 시스템을 나타낸다.

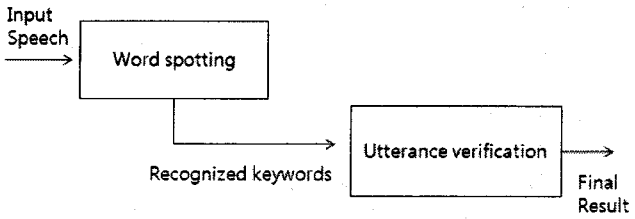

그림 2. 핵심어 인식 및 발화검증 시스템
Figure 2. Word spotting system using an utterance verification system

2.2 LRT 를 기반으로 한 발화검증

발화검증은 가설검정의 과정이다. 음성 특징 벡터 X에 대해 음성인식 시스템이 이를 λ_w라는 HMM 모델로 표현되는 단어로 인식한다고 하자. 다음과 같이 상호 보완하는 두 가지 가설을 제안한다.

H_0: 영가설 (null hypothesis), X 가 올바르게 인식되었고 모델 λ_w에 포함됨

H_1: 대안가설 (alternative hypothesis), X 가 틀리게 인식되었고 모델 λ_w에 포함되지 않음

Pearson Lemma 에 의하면 영가설 (null hypothesis)과 대안가설 (alternative hypothesis) 하에서의 우도를 정확히 알 수 있을 경우 LRT가 가설검정에 가장 적합한 해결방안을 제시한다.

$$LR = \frac{P(X|H_0)}{P(X|H_1)} \tag{1}$$

위의 비율값을 임계치와 비교하여 수락 / 거절을 결정한다. 올바르게 디코딩된 경우를 나타내는 영가설(null hyothesis) 하에서의 우도와 틀리게 디코딩된 경우를 나타내는 대안가설(alternative hypothesis) 하에서의 우도가 주어졌을 때 LRT는 가정된 결과를 수락 또는 거절하는 테스트를 나타낸다.

LRT의 가장 어려운 점은 대안가설을 어떻게 모델링하느냐에 있다. 본 논문에서는 각 triphone 단위로 인식된 음소의 경계 내에서 그와는 다른 음소에 해당하는 monophone들에 대한 우도값을 구한다. 이와 같이 구한 값 중 가장 큰 우도 값을 갖는 monophone을 대안가설에 대한 모델, 즉 반 모델로 선택한다. Cohort 모델을 사용해 반 모델을 구성하는 경우, 인식된 음소에 대해 그 음소에 해당하는 cohort 모델로 항상 동일한 음소들이 반 모델로 선정되는 반면, 이 경우 같은 음소에 대해서도 음의 특성에 따라 다른 음소가 반 모델로 선정될 수 있다는 특성을 지닌다.

이러한 방식은 인식과정에서의 경쟁적 정보에 기인한다. 인식된 결과가 올바르다면 인식된 모델은 경쟁하는 모델들과 월등한 차이를 보일 것이며 따라서 영가설 하에서의 우도가 대안가설 하에서의 우도에 비해 큰 값을 갖게 될 것이다. 이에 따라 LRT 비율값이 크게 될 것이다. 반면 인식된 결과가 틀리다면 인식된 모델과 경쟁하는 모델이 큰 차이를 보이지 않을 것이며 따라서 LRT 비율값이 작게 될 것이다. 이와 같이 본 논문에서 사용된 방식은 triphone 단위의 인식된 결과가 경쟁자들과 얼마나 큰 차이를 보이느냐에 따른 비율값을 구하여 이 값에 따른 수락 혹은 거절하는 이론에 기초하고 있다. <그림3>에 각 단어에 대한 신뢰도 척도를 구하는 과정이 나타나 있다.

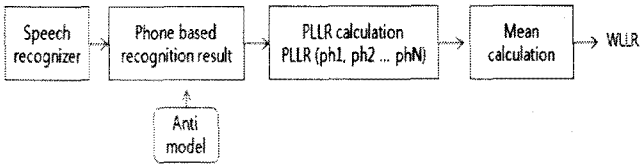

그림 3. 각 단어에 대한 신뢰도 척도를 구하는 과정
Figure 3. Procedure for generating confidence measures for each word

<그림3>은 LRT 기반의 기본적인 방식으로, 각 PLLR 값을 구해 이 값들의 평균을 취해 word-level log-likelihood ratio (WLLR)을 구하는 방식을 나타낸다. PLLR 은 LRT 를 음소단위로 적용하여 다음과 같이 구한다.

$$PLLR(\text{ph}) = \frac{\log P(X_{\text{ph}} \mid \lambda_{\text{ph}}) - \log P(X_{\text{ph}} \mid \bar{\lambda}_{\text{ph}})}{\tau(\text{ph})} \tag{2}$$

여기서 X 는 인식된 음소의 입력 특징벡터, τ 는 인식된 음소 ph 의 지속시간, λ 와 $\bar{\lambda}$ 는 각각 인식된 모델에 대한 음향모델 그리고 반 모델을 나타낸다. WLLR 로 나타내어지는 각 단어에 대한 신뢰도는 다음과 같이 구한다.

$$C_{ari}(w) = \frac{1}{n_p(w)} \sum_{j=1}^{n_p(w)} sigmoid(PLLR(\text{ph}_j)) \tag{3}$$

여기서 $n_p(w)$ 는 단어 w 를 구성하는 음소 수를 나타내고 sigmoid 함수는 다음과 같다.

$$sigmoid(x) = \frac{1}{1+\exp(-ax-b)} \tag{4}$$

a와 b는 실험적으로 구한다. 본 논문에서는 식 (3)의 c_{ari} 값을 가장 기본적인 LRT 기반의 신뢰도 척도로 사용하였다.

본 논문에서는 핵심어 인식 시스템의 결과에 대한 발화검증을 수행하는 시스템을 구현하였다. 일반적인 단어 인식시스템과 핵심어 인식 시스템에 대한 발화검증의 가장 큰 차이점은 일반적인 단어 인식 시스템이 입력음성 전체에 대한 검증을 하는 반면 핵심어 인식 시스템은 핵심어로 인식된 음성의 특정 부분에 대해서만 검증을 한다는 점이다. 특히 인식된 핵심어는 단어 인식기의 인식 결과에 비해 단어 경계가 부정확하다는 특성을 지니며 이에 따라 두 시스템에 대한 발화검증은 조금 다른 특성을 갖게 된다.

2.3 발화검증을 위한 Word Voiceprint 방식

본 절에서는 기존의 방식 중 단어 내의 각 음소의 PLLR 분

포를 사용하는 Word Voiceprint 방식 [5]을 소개한다. Word Voiceprint는 다음과 같은 관찰로부터 제안되었다. 실험결과 올바른 단어와 오인식된 단어의 구성음소의 차이가 크지 않은 경우 낮은 PLLR 값을 갖는 몇 개의 음소에도 불구하고 다른 음소의 PLLR 값들의 영향으로 WLLR 이 임계치보다 크게 되어 단어를 수락하게 되는 경우가 발생하였다. 또한 PLLR 값들을 분석해본 결과 같은 음소에 대해서도 단어가 다른 경우 다른 PLLR 분포를 가지고 있음을 볼 수 있었다. Word Voiceprint는 각 단어의 음소별 PLLR 분포를 사용하여 각 단어에 보다 적합한 신뢰도를 구하는 방식이다. Word Voiceprint 에서는 음소단위의 신뢰도 척도를 다음과 같이 구한다.

$$C_{VP}(ph)=\begin{cases}0, & PLLR(ph)\geq \mu_w^{ph}-\sigma_w^{ph}\\ \psi(PLLR(ph)), & PLLR(ph)<\mu_w^{ph}-\sigma_w^{ph}\end{cases} \tag{5}$$

μ_w^{ph} 와 σ_w^{ph} 는 각각 단어 w 에서 음소 ph 의 PLLR 값의 평균과 표준편차를 의미한다.

$\psi(\bullet)$는 다음과 같이 구한다.

$$\psi(PLLR(ph))=\log\frac{e^{\beta}(PLLR(ph)-(\mu_w^{ph}-\alpha\sigma_w^{ph}))}{(\alpha-1)}-\beta \tag{6}$$

α와 β는 실험적으로 구한 값을 사용하였다. <그림4>는 $\psi(\bullet)$ 함수를 보여준다. 어떤 단어에서 한 음소라도 PLLR 값이 $\mu_w^{ph}-\alpha\sigma_w^{ph}$ 보다 떨어질 경우 그 단어는 거절될 가능성이 높게 된다.

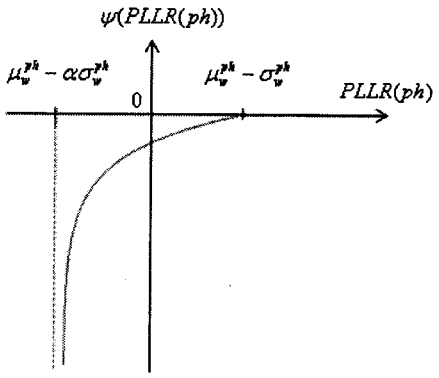

그림 4. 음소단위의 voiceprint 신뢰도 척도를 구하기 위한 로그 스케일의 비 선형적 함수

Figure 4. Log-scale non-linear function to obtain phone-level voiceprint confidence score

최종적으로 단어 단위의 신뢰도 척도를 다음과 같이 각 음소 단위의 신뢰도척도의 산술평균으로 구한다.

$$C_{VP}(w)=\frac{1}{n_p(w)}\sum_{j=1}^{n_p(w)}C_{VP}(ph_j) \tag{7}$$

Word Voiceprint는 다른 단어에서 각 음소의 PLLR 값들의 분포를 이용하여 올바른 단어와 오인식된 단어 사이의 구성음소의 차이가 크지 않은 경우 단어를 더 효과적으로 거절한다.

3. 음소레벨 로그우도 패턴을 이용한 발화검증

3.1 기존 LRT 방식의 문제점

<그림5>에 나타난 예는 올바르게 인식되었으나 정확한 발성과는 약간의 차이를 갖고 발성된 카드번호라는 단어에 대한 예이다. 정확하게 발성되지 않은 n 음소에 의해 n 에서의 PLLR 값이 낮게 나온 특성을 보였다. 또한 부정확한 단어 경계에 의하여 첫 번째 음소와 마지막 음소가 낮게 나오게 되었다. 올바르게 인식된 핵심어임에도 불구하고 이러한 낮은 PLLR 값들의 영향으로 결과적인 WLLR 값을 임계치와 비교하였을 때 발화검증시스템은 단어를 거절하게 되는 현상이 발생하게 되었다.

각 단어의 다양한 발화에 대해 이와 같은 PLLR 값들을 관찰해본 결과 각 음소마다 이처럼 PLLR 값이 떨어질 확률이 다르다는 것을 볼 수 있었다. 본 논문에서는 이와 같이 PLLR 값이 떨어질 가능성이 높은 음소에 대한 방안으로 각 단어의 음소마다 다른 가중치를 부여하는 방식을 제안한다. 제안한 방식에 따

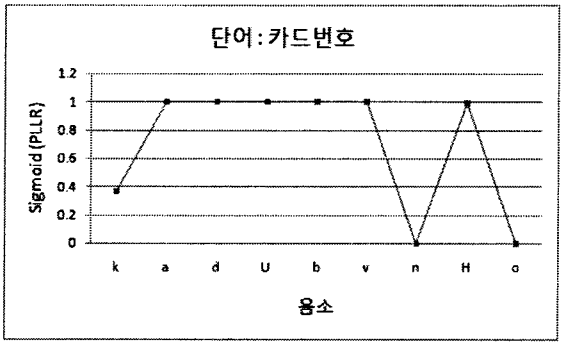

그림 5. 단어의 음소에 따른 PLLR 값

Fig 5. PLLR values of a correctly recognized keyword

라 가중치를 생성하며 이와 같은 가중치는 단어에 대한 신뢰도 척도를 구하는 과정에서 각 단어의 음소에 대해 다른 가중치를 부여한다.

3.2 제안된 신뢰도 척도를 구하는 방법

각 음소에 대한 가중치를 생성하기 위해서는 일단 각 핵심어별 다양한 발화에 대해 PLLR 의 패턴을 분석한다. 특정 발화의 어떤 음소에서 PLLR 값이 그 발화의 PLLR 값의 평균보다 떨어지게 될 경우, 그 음소에 대해 PLLR distance라는 값을 부여

한다. <그림6>에서 PLLR distance의 예를 보인다. 이 경우에는 3번째 음소의 PLLR 값이 그 발화음성에서 PLLR 값의 평균에 비해 떨어진 경우이며 이 발화 단어의 3번째 음소에 대해 PLLR distance가 부여되었다. PLLR distance는 단어에 대한 특정발화에서 PLLR 값이 다른 음소에 비해 얼마나 떨어지는지를 나타낸다. 이와 같은 PLLR distance를 각 핵심어 별로 구해 각 핵심어의 음소별 가중치를 생성하는 기본적인 요소로 사용한다.

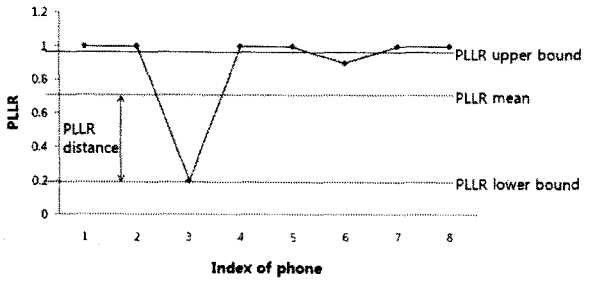

그림 6. PLLR distance 값의 예
Figure 6. An example of a PLLR distance

제안된 방식에서는 다음과 같이 각 음소에 곱해지는 가중치를 구한다.

$$d(ph^i_{j,w}) = \begin{cases} \exp\left(\alpha\left(\dfrac{\mu^i_w - PLLR(ph^i_{j,w})}{ub^i_w - lb^i_w}\right)\right) - 1, & PLLR(ph^i_{j,w}) < \mu^i_w \\ 0, & otherwise \end{cases} \tag{8}$$

여기서 i는 발화에 대한 index를 나타내고, j는 음소에 대한 index, w는 인식대상으로 포함된 핵심어, $ph^i_{j,w}$는 핵심어 w 에 대한 i번째 발화의 j번째 음소를 나타낸다. μ^i_w, ub^i_w 및 lb^i_w 는 각각 핵심어 w의 i번째 발화에서 PLLR 값의 평균, 상한, 하한 값을 나타낸다.

정규화된 PLLR distance를 비선형적으로 구한 $d(ph^i_{j,w})$는 핵심어에 대한 발화에서 특정음소에서의 PLLR 값이 다른 음소들에 비해 얼마나 낮은가를 나타내는 값이다. α는 핵심어에 따라 다른 값을 사용하였다. 핵심어의 각 음소에 대한 $\bar{d}(ph_{j,w})$는 다음과 같이 구한다.

$$\bar{d}(ph_{j,w}) = \frac{1}{N_w}\sum_{i=1}^{N_w} d(ph^i_{j,w}) \tag{9}$$

N_w는 평가셋에 해당하는 데이터에서 핵심어 w에 대한 발화수이다. $d(ph^i_{jw})$값의 평균값인 $\bar{d}(ph_{j,w})$는 핵심어의 각 음소별로 PLLR 값이 떨어질 가능성을 나타내는 값이다. $\bar{d}(ph_{j,w})$ 값이 큰 음소에 대해서 적은 가중치를 부여 하는 것을 목적으로 하기 때문에 다음과 같은 감소함수를 이용하여 핵심어의 각 음소에 대한 가중치를 결정한다.

$$\tilde{c}(ph_{j,w}) = \exp(-\gamma \cdot (\bar{d}(ph_{j,w}) - \tau)) \tag{10}$$

γ, τ는 계수들간의 편차를 조정한다. 임의적으로 구한계수를 다음 식에 의해 0에서 1 사이의 확률적인 계수 값으로 변환한다.

$$c(ph_{j,w}) = \frac{\tilde{c}(ph_{j,w})}{\sum_{j=1}^{n_p(w)} \tilde{c}(ph_{j,w})} \tag{11}$$

$n_p(w)$는 각 핵심어에 대한 음소 수이다. 최종적으로 다음과 같이 PLLR 패턴을 이용한 단어별 신뢰도를 구한다.

$$CM = \frac{1}{n_p(w)}\sum_{j=1}^{n_p(w)} c(ph_{j,w}) \cdot sigmoid(PLLR(ph_{j,w})) \tag{12}$$

(8)에서 보면 비 선형적인 PLLR distance 값을 구하여 가중치를 구하는 과정에서 사용한 것을 볼 수 있다. 비 선형적인 PLLR distance를 사용한 이유는 실험을 수행해본 결과 선형적인 PLLR distance를 사용하여 가중치에 대해 영향을 준 경우에 비해 비 선형적인 PLLR distance를 사용해 PLLR distance가 낮을 값을 나타냈을 때의 영향을 가중치의 형성에 더 확실히 반영하였을 때 더 효과적이었기 때문이다. 실제 실험에서는 음소에서의 PLLR 값이 발화의 평균에 비해 낮은 값을 나타낼때 $d(ph^i_{j,w})$ 대신 $d(ph^i_{j,w})$ +1을 사용하여 그 음소의 가중치에 대한 영향을 더 확실히 주었다. 큰 PLLR distance 값들을 나타내는 음소들에 대해 더 작은 가중치를 부여하도록 하기 때문에 감소함수를 사용하고 (10)에 나타난 감소함수를 사용하였다. α, γ 및 τ는 개발 셋에 대한 실험을 통하여 구하였다. 감소함수의 γ, τ 값을 사용해 계수들간의 편차를 조정하게 하였지만 실제 실험에서는 γ, τ 에 의한 영향은 크지 않았다.

4. 실험 및 결과

4.1 실험조건

본 논문에서는 한국어 전화망 환경 음성인식용 대화체 문장 데이터베이스를 사용하였다. 데이터베이스에 대한 구체적인 설명을 <표1>에서 나타낸다.

표 1. 음성데이터베이스에 대한 설명
Table 1. Description of speech database

구분	설명	부가설명
데이터베이스 명칭	한국어 전화망 환경 음성인식용 대화체 문장 DB	
발화형태	연속어	
발화방식	대화체	
녹음방식	전화상담원과 상담자간의1:1대화 음성을 녹음	렌트카예약,호텔예약, 영화예매 등의 시나리오에 대해 녹음
화자 수	10~20명의 상담원 화자	
녹음 상태	-다양한 잡음환경 -15dB이상의 SNR	가정, 사무실, 거리등의 환경
음성 데이터 수	5,500개의 음성데이터	음성데이터 당 3분정도의 길이

각 음성 데이터는 렌트카예약, 관광문의, 호텔예약 등의 시나리오 영역에서 상담원과 고객간의 대화로 구성되어 있다. 잡음환경하에서 녹음되었으며 총 5,500개의 음성 데이터로 이루어진다. 실제 실험에서는 음성데이터를 훈련, 개발 평가 셋으로 3등분하여 사용하였다.

훈련 셋은 핵심어 및 비핵심어 모델을 훈련할 목적으로 사용된다. 전체 데이터 중 4,500개의 음성데이터를 사용하였고 4,500개의 음성 데이터에서 핵심어에 해당하는 부분을 수작업으로 분할한 음성을 사용하여 핵심어 모델을 훈련하였다. garbage 모델은 훈련 셋 전체의 4,500개의 음성을 사용하여 훈련하였다. 개발 셋은 각 핵심어의 PLLR 패턴을 분석할 목적으로 사용된다. 전체 데이터 중 500개의 음성데이터를 사용하였고 500개의 음성데이터에 대해 핵심어 인식을 하여 인식된 총 819개의 핵심어로 각 핵심어별 PLLR 패턴정보를 분석하였다. 평가 셋은 핵심어 인식 시스템과 제안된 발화검증 시스템의 성능평가를 위해 사용된다. 전체 데이터 셋에서 훈련 및 개발 셋과는 별개의 500개의 음성데이터를 사용하였다. 평가 셋에는 실험에 사용된 5개의 핵심어를 기준으로 하였을 때 총 911개의 핵심어가 존재한다. 평가 셋에 해당하는 각 음성 데이터에는 2~3개의 핵심어가 존재한다. 각 핵심어 및 garbage 모델은 훈련 셋을 기반으로 훈련되었으며, 상세한 내용은 <표2>와 같다.

표 2. 핵심어 및 garbage모델에 대한 설명
Table 2. Description of model training

모델 / 설명	핵심어	비 핵심어
훈련에 사용된 음성	4,117개의 핵심어 단위로 분할된 음성	전체 4,500개의 훈련 셋에 해당하는 음성
HMM	3 tied-states, 7개의 가우시안 분포를 갖는 HMM	3 states 5개의 가우시안 분포를 갖는 HMM
모델링 단위	triphone	monophone
훈련에 사용된 특징 벡터	12MFCC + Energy의 13차 기본벡터 + delta 13차 + delta-delta 13차	12MFCC+Energy의 13차 기본벡터 +delta 13차 +delta-delta 13차

4.2 발화검증에 대한 실험결과

각 방식의 성능평가를 위해 다음과 같은 실험을 수행하였다. 제안된 방식에 대한 실험, Word Voiceprint 방식을 사용한 실험 그리고 제안된 방식과 Word Voiceprint를 결합하여 사용한 경우에 대한 실험을 수행하였다. 성능 평가는 올바르게 인식된 핵심어가 거절되는 비율을 나타내는 false rejection rate 오인식된 핵심어가 수락된 비율을 나타내는 false acceptance rate 이 같은 경우를 나타내는 equal error rate(EER) 및 오류(error)가 감소한 비율을 나타내는 error reduction rate(ERR)을 사용하였다. <표3>에 각 방식에 의한 실험결과를 나타낸다.

표 3. EER (%) 로 평가된 제안된 방식의 성능
Table 3. Performance of proposed method in EER (%)

방식 / 핵심어	기본적인방식	제안된 방식	ERR (%)
운전면허증	14.0	11.5	17.9
영화시작	9.5	8.9	6.3
요금제	24.1	24.3	0
신용카드	16.9	12.2	27.8
카드번호	7.1	6.6	7.0
overall			11.8

기본적인 방식은 각 핵심어에 대해 기본적인 LRT 방식을 사용해 각 단어에 대한 신뢰도 척도를 구하는 방식을 의미한다. 제안된 방식은 PLLR 패턴정보를 이용하여 각 단어에 대한 신뢰도 척도를 구하는 방식을 의미한다.

실험결과 핵심어 '요금제'를 제외한 모든 핵심어에서 제안된 방식에 의한 성능향상 이 있는 것을 볼 수 있었다. 핵심어 '요금제'는 적은 수의 음소를 갖고 있었으며 이 핵심어는 뚜렷한 PLLR 패턴을 보이지 않았다. 전체적인 ERR을 분석해 본 결과 제안된 방식을 사용하였을 때 기본적인 방식에 비해 약 11.8%의 성능향상을 나타내는 것을 볼 수 있었다. <표4>는 Word Voiceprint 방식, 제안된 방식과 Word Voiceprint를 결합한 방식에 대하여 실험을 수행하였을 때의 성능을 나타낸다.

표 4. EER (%) 로 평가된 Word Voiceprint와 제안된 방식과 Word Voiceprint를 결합한 방식의 성능
Table 4. Performance of word voiceprint method and combined method in EER (%)

방식 / 단어	기본적인 방식	Word Voiceprint	제안된 방식	제안된 방식 +Word Voiceprint
운전면허증	14.0	9.0	11.5	7.0
영화시작	9.5	9.9	8.9	7.9
요금제	24.1	19.7	24.3	20.9
신용카드	16.9	13.3	12.2	12.0
카드번호	7.1	6.4	6.6	6.8
ERR (%)		16.2	11.8	22.6

Word Voiceprint는 각 핵심어에 대한 신뢰도 척도를 구하기 위해 Word Voiceprint 발화검증 방식을 사용한 경우를 나타낸다. 제안된 방식 + Word Voiceprint는 제안된 방식과 Word Voiceprint 방식을 결합한 발화검증 방식을 사용하여 각 핵심어에 대한 신뢰도 척도를 구한 경우를 나타낸다. 이 경우 제안된 방식에 의한 계수들은 각 Word Voiceprint 기반의 신뢰도 척도에 가중치를 부여한다. 실험결과를 분석해 보면 Word Voiceprint를 사용한 경우 성능의 향상이 있는 것을 볼 수 있었다. Word Voiceprint를 사용한 경우 핵심어 '요금제'에 대해서도 성능향상을 가져올 수 있었다. 이러한 결과는 각 핵심어의 PLLR 분포를 이용하여 각 핵심어에 대한 신뢰도 척도를 구한 경우 단어의 특성에 보다 적합한 신뢰도를 구할 수 있다는 것에 기인한다.

전체적인 ERR을 분석해 보면 Word Voiceprint 방식을 사용한 경우 기본적인 방식에 비해 16.2%의 성능향상을 나타내어 기본적인 방식에 비해 더 좋은 성능을 나타내는 것을 볼 수 있었다. 제안된 방식과 Word Voiceprint 방식을 결합한 경우 추가적인 성능향상을 보였고 결과적으로 22.6%의 성능향상을 보이고 있다.

4.3 핵심어 인식에 대한 실험결과

<표5>는 기본적인 방식, 제안된 방식 그리고 Word Voiceprint 방식을 사용하여 인식된 핵심어에 대한 발화검증을 수행한 핵심어 인식시스템의 성능을 나타낸다. 핵심어 인식 시스템에 대한 성능평가는 올바르게 인식된 핵심어의 개수를 평가 셋에 해당하는 음성 내에 존재하는 핵심어의 개수로 나눈 핵심어 검출율, false accept 된 핵심어의 개수를 인식대상 핵심어의 개수와 평가 셋에 해당하는 음성의 총 시간을 곱한 값으로 나눈 FA/keyword/hour를 사용하였다.

표 5. 각 발화검증 방식을 사용한 핵심어 인식 시스템의 성능
Table 5. Word spotting performance using various utterance verification methods

Word insertion penalty	핵심어 검출율(%)	오인식률 (FA/keyword/hour)		
		기본적인 방식	제안된 방식	제안된 방식 +Word Voiceprint
-10	72.33	0.23	0.20	0.19
-20	74.75	0.42	0.40	0.33
-50	85.84	1.62	1.28	1.21
-70	87.71	3.94	2.36	2.06
-90	87.26	6.85	4.6	3.98

기본적인 방식은 기본적인 방식의 발화검증을 사용하여 핵심어 인식을 수행한 결과를 나타낸다. 제안된 방식은 제안된 방식의 발화검증을 사용하여 핵심어 인식을 수행한 결과를 나타내고 제안된 방식 + Word Voiceprint 는 제안된 방식과 Word Voiceprint 방식을 결합하여 발화검증을 수행한 결과를 나타낸다.

각 핵심어 모델과 garbage 모델에는 동일한 word insertion penalty를 부여하였다. Word insertion penalty 값을 줄일 경우 더 많은 핵심어를 검출할 수 있는 반면 오인식된 핵심어가 증가한다는 단점을 갖는다. Word insetion penalty 값이 -70일때 성능이 포화된 것을 볼 수 있다. 핵심어 인식 결과는 각 word insertion penalty에서 각 방식 당 총 인식된 핵심어에서 5%의 false rejection rate을 갖도록 발화검증을 수행하여 얻게 된 결과이다. 따라서 성능비교를 위해서는 FA/keyword/hour 값을 비교한다. 각 word insertion penalty 값에서 동일한 핵심어 검출율 하에서 제안된 방식은 기본적인 방식에 비해 낮은 FA/keyword/hour 값을 나타내어 성능향상을 보인다. 또한 제안된 방식과 Word Voiceprint 방식을 결합하여 사용한 경우 기존의 방식과 제안된 방식에 비해 추가적인 성능향상을 가져올 수 있었다.

5. 결 론

본 논문에서는 핵심어 인식 시스템과 핵심어 인식시스템의 후단에 사용되는 발화검증 시스템을 구현하였다. 기본적인 LRT 기반의 발화검증을 사용하여 핵심어 인식 시스템으로부터 인식된 핵심어에 대한 발화검증을 수행하였을 때 몇 가지 문제점이 있었다. 이러한 문제점들로 인하여 올바르게 인식된 핵심어임에도 불구하고 몇 개의 음소가 낮은 PLLR 값을 보여 단어에 대한 신뢰도 척도, 즉 WLLR 이 낮게 나와 핵심어를 거절하게 되는 경우가 발생하였다. 이와 같은 문제점을 해결하기 위해 본 논문에서는 단어에 대한 신뢰도 척도를 구할 때 각 음소별로 다른 가중치를 부여하는 방식을 제안하였다. 또한 단어의 각 음소별 PLLR 의 분포를 사용하는 Word Voiceprint 를 검토해 보았다. 실험결과, 제안된 방식에 의해 성능향상을 보일 수 있었다. 또한 Word Voiceprint 방식을 사용한 경우 추가적인 성능향상을 가져올 수 있었다.

제안된 방식은 모든 인식대상 핵심어에 대한 패턴분석을 해야 한다는 단점을 지닌다. 이러한 단점을 극복하기 위해 단어에 대한 패턴을 얻기 위해 보다 간단한 방식이 필요함을 볼 수 있었다. 보다 간단한 방식에 대한 연구를 위해 앞으로는 각 음소들을 특성에 따라 군집화하여 핵심어의 변동이 있을 경우 군집화된 정보를 사용하여 핵심어에 대한 패턴 정보를 더 용이하게 구할 수 있는 방식 등에 대한 연구를 수행할 것이다.

참고문헌

[1] A. L. Higgins and R. E. Wohlford, "Keyword recognition using template concatenation", *IEEE International Conf. on Acoustics, Speech, and Signal processing (ICASSP 85)*, Vol. 10, pp. 1233-1236, 1985.

[2] R. C. Rose and D.B. Paul, "A hidden markov model based keyword recognition system", *IEEE International Conf. on*

Acoustics, Speech, and Signal Processing (ICASSP 90), Vol. 1, pp. 123-132, 1990.

[3] P. Heracleus and T. Shimizu, "A novel approach for modeling non-keyword intervals in a keyword spotter exploiting acoustic similarities of languages", *Speech Communication*, Vol. 45, pp. 373-386, 2005.

[4] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 4, pp. 420-429, 1996.

[5] S. -B. Kwon and H. -R. Kim, "Utterance verification using word voiceprint models based on probabilistic distributions of phone-level log-likelihood ratio and phone duration", *IEICE Transactions on Information and Systems*, Vol. E91-D, No. 11, pp. 2746-2750, 2008.

[6] K. -S. Moon, Y. -J. Kim, H. -R. Kim and J. -H. Chung, "Out-of-vocabulary word rejection algorithm in korean variable vocabulary word recognition", *IEEE International Symposium on Circuits and Systems*, Vol. 5, pp. 53-56, 2000.

• **김정현 (Kim, Chong-Hyon)**
주소: 305-732 대전광역시 유성구 문지동 103-6
한국과학기술원 ICC
Tel: 042-350-6221
Email: polarbear@icu.ac.kr
관심분야: 핵심어 인식, 발화검증
현재: 한국과학기술원 정보통신공학과 박사과정

• **권석봉 (Kwon, Suk-Bong)**
주소: 305-732 대전광역시 유성구 문지동 103-6
한국과학기술원 ICC
Tel: 042-350-6221
Email: sbkwon@icu.ac.kr
관심분야: 음성인식 탐색 알고리즘, 발화검증
현재: 한국과학기술원 정보통신공학과 박사과정

• **김회린 (Kim, Hoi-Rin)**
주소: 305-732 대전광역시 유성구 문지동 103-6
한국과학기술원 ICC
Tel: 042-350-6139
Email: hrkim@ee.kaist.ac.kr
관심분야: 음성인식, 핵심어 인식, 화자인식, 오디오신호처리
현재: 한국과학기술원 정보통신공학과 부교수