다채널 마이크 환경에서 Naive Bayesian Network의 Decision에 의한 음성인식 성능향상

지미경, 김회린
한국정보통신대학교 공학부 음성인식기술연구실

Performance Improvement in Distant-Talking Speech Recognition by an Integration of *N*-best results using Naive Bayesian Network

Mikyong Ji, Hoi-Rin Kim
School of Engineering, Information and Communications University

{lindaji, hrkim}@icu.ac.kr

Abstract

원거리 음성인식에서 인식률의 성능향상을 위해 필수적인 다채널 마이크 환경에서 방 안의 도처에 분산되어 있는 원거리 마이크를 사용하여 TV, 조명 등의 주변 환경을 음성으로 제어하고자 한다. 이를 위해 각 채널의 인식결과를 통합하여 최적의 결과를 얻고자 채널의 *N*-best 결과와 *N*-best 결과에 포함된 hypothesis의 frame-normalized likelihood 값을 사용하여 Bayesian network을 훈련하고 인식결과를 통합하여 최선의 결과를 decision 하는데 사용함으로써 원거리 음성인식의 성능을 향상시키고 또한 hands-free 응용을 현실화하기 위한 방향을 제시한다.

Ⅰ. 서론

현재 음성인식 시스템은 가까운 거리에서 사용 시, 높은 인식률을 보여주고 있지만 발성 화자가 마이크에서 거리가 멀어질수록 반향 효과, 배경 잡음 또는 신호 자체의 SNR 저하 등의 이유로 인식률이 급격히 떨어지는 것을 볼 수 있다. 최근 원거리 음성인식 연구 동향에서는 전처리 모듈로서 어레이 마이크 시스템을 사용하는 방법이 제안되었고 보다 효율적인 성능향상을 보여주고 있으며, hands-free 음성인식 분야에 대한 희망적인 해답을 제시하고 있다[1, 2, 3]. 그럼에도 불구하고 어레이 마이크 시스템 사용 시, talker location 과정에서의 에러가 오히려 음성인식의 성능저하를 초래하고 있다[4]. 유비쿼터스 환경에서 hands-free 응용을 보다 현실적으로 하기 위해서는 핵심 기술인 원거리 음성인식과 다채널 마이크 입력에 의한 얻어진 인식결과 중 최적의 결과를 선택하는 즉 인식결과를 통합하는 하는 기술이 필요하다[5].

본 논문에서는 다채널 음성입력에 의한 likelihood score 값에 기반을 둔 통합 방법과 Naive Bayesian network을 이용하여 다채널 인식결과를 통합하는 방법 두 가지를 제시하고 그 성능을 비교해 본다. 통합 방법에 있어서, 각 채널로부터의 입력음성은 모델이 주어졌을 때 서로 독립적인 관계에 있다고 가정하였다. 각 채널로부터의 인식결과를 통합하기 위한 전체 시스템 구조는 그림 3과 같다.

2장에서는 다채널 마이크에 의한 N-best 인식결과를 통합하기 위한 두 가지 방법 likelihood 기반의 통합방법과 Bayesian network에 의한 통합에 대해 자세히 설명하고 3장에서는 두 가지 통합 방법의 성능을 비교하고 평가한다.

본 연구는 정보통신부 및 정보통신연구진흥원의 디지털미디어 연구소 지원사업의 연구결과로 수행되었음

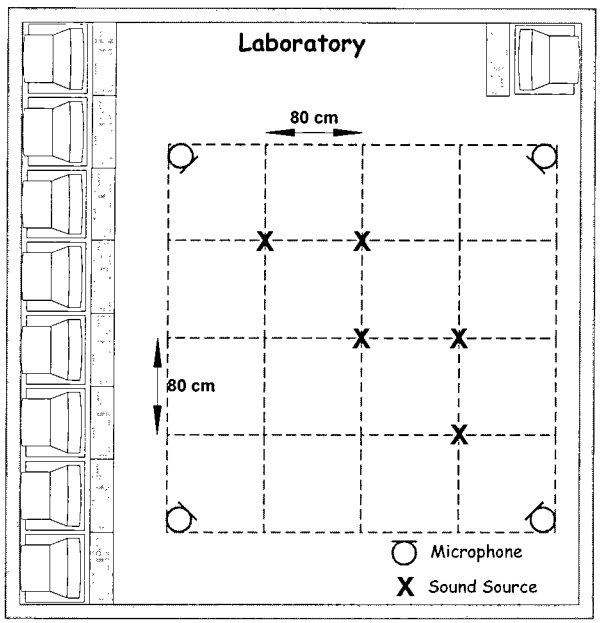

그림 1. 분산된 원거리 마이크의 위치 및 실험환경

Ⅱ. N-best 결과의 Integration

1. Likelihood 기반의 Integration Method

방안의 도처에 분산되어 있는 각 마이크에 의해 입력되는 음성은 모델이 주어졌을 때 서로 독립적이라 가정하면, 식 (1)과 (2)에서와 같이 단어 별 채널의 입력음성에 대한 posterior probability의 합에 의해 인식결과를 결정할 수 있다.

$$\begin{aligned}\overline{W} &= \arg\max_{W} P(W|X_1,X_2,X_3,X_4) \qquad (1)\\ &= \arg\max_{W} \frac{P(W)\prod_{i=1}^{4}P(X_i|W)}{\sum_{W}P(W)\prod_{i=1}^{4}P(X_i|W)}\\ &= \arg\max_{W} \frac{\prod_{i=1}^{4}P(X_i|W)}{\sum_{W}\prod_{i=1}^{4}P(X_i|W)}\\ &= \arg\max_{W} \frac{\prod_{i=1}^{4}P(W|X_i)}{\sum_{W}\prod_{i=1}^{4}P(W|X_i)}\\ &\cong \arg\max_{W} \prod_{i=1}^{4}P(W|X_i)\end{aligned}$$

X_i 는 채널 i로부터 들어오는 음성입력을 나타내고, W 는 각 단어모델을 나타낸다. 즉 채널 별 음성입력이 주어졌을 때, $P(W|X_1,X_2,X_3,X_4)$을 최대로 하는 단어를 인식결과로 결정한다. 각 채널입력 X_i는 W가 주어지면 서로 독립적이다 라는 가정에 의해 식 (1)과 (2)와 같이 전개할 수 있다. 본 논문에서는 식 (2)에 의한 단어 별 각 채널입력에 따른 posterior probability의 합에 의해 인식단어를 결정하는 방법의 인식성능을 측정하였다.

$$\begin{aligned}\overline{W} &= \arg\max_{W} \prod_{i=1}^{4}P(W|X_i) \qquad (2)\\ &= \arg\max_{W} \prod_{i=1}^{4}\frac{P(X_i|W)}{\sum_{W\in N-best}P(X_i|W)}\\ &= \arg\max_{W} \sum_{i=1}^{4}\log\frac{P(X_i|W)}{\sum_{W\in N-best}P(X_i|W)}\end{aligned}$$

2. Naive Bayesian Network 기반의 Integration Method

Naive Bayesian Network(BN) 기반의 integration 방법의 경우, N-best 결과의 likelihood 값 외에 순서정보를 사용하여 Bayesian network을 구성하고 이를 훈련하여 다채널 원거리 환경에서의 음성인식 성능을 향상시킨다. 그림 2에서 보는 바와 같이 단어 별 각 채널로부터의 N-best list에서의 순서정보와 앞서 설명했던 likelihood 기반의 integration에 의한 인식결과를 이용하여 BN을 구성하고 채널의 인식결과를 통합하는데 사용하였다.

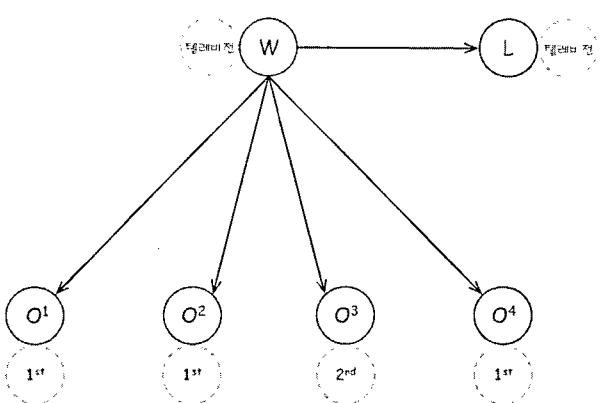

그림 2. 각 채널에 의한 N-best 결과를 통합하여 최선의 결과를 결정하기 위한 Naive Bayesian Network

Oi는 채널 i의 음성입력에 의한 N-best 결과에서 단어 W의 순서정보를 나타내며, L은 앞서 설명했던 likelihood 기반의 integration 방법에 의한 인식결과를 나타내고 W는 인식 어휘를 나타낸다. 그림 2에서의 BN 구조가 나타내 듯 인식단어 W가 주어지면 Oi는

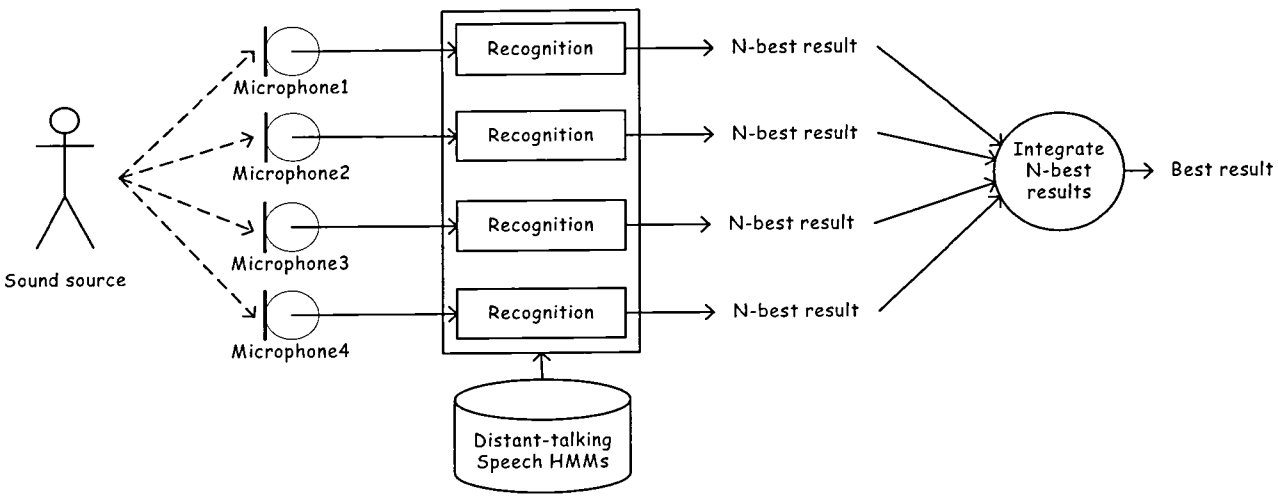

그림 3. 다채널 음성입력에 의한 N-best 결과 통합을 위한 전체 시스템 구조

서로 독립적이다. 각 Oi는 1부터 N까지의 값을 가질 수 있고, W와 L 은 총 어휘 수인 30까지의 값을 가질 수 있다. 그림 2에 점선의 노드 안에 각 변수가 가질 수 있는 값의 예를 표시하였다.

음성이 각 마이크를 통해 입력되면 각 채널 입력에 대한 N-best 인식결과를 얻을 수 있다. N-best 인식결과 내에 포함된 모든 단어 W에 대해 각 채널에서의 N-best list 내에 순서와 likelihood 기반의 integration에 의한 인식결과가 주어졌을 때, W가 정답일 확률을 최대로 하는 W가 인식결과로 결정한다. 따라서 $P(W|O1,O2,O3,O4,L)$은 Markov condition에 의해 식 (3)과 같이 전개할 수 있다. Markov condition 이란 BN에서 각 node는 node의 parent가 주어지면 모든 non-descendant node와 독립적인 관계에 있다는 조건이다. 수식으로 표현하면 식 (4)와 같다.

$$\overline{W} = \underset{W \in N-best}{\arg max} P(W|O^1,O^2,O^3,O^4,L) \tag{3}$$

$$= \underset{W \in N-best}{\arg max} \frac{P(W,O^1,O^2,O^3,O^4,L)}{\sum_W P(W,O^1,O^2,O^3,O^4,L)}$$

$$= \underset{W \in N-best}{\arg max} \frac{P(W)P(L|W)\prod_{i=1}^{4} P(O^i|W)}{\sum_W P(W)P(L|W)\prod_{i=1}^{4} P(O^i|W)}$$

$$\cong \underset{W \in N-best}{\arg max} P(W)P(L|W)\prod_{i=1}^{4} P(Oi|W)$$

$$I_P(X, ND_X | PA_X) \tag{4}$$

$P(W)$, $P(O^i|W)$와 $P(L|W)$는 각 node의 parent와의 의존성을 고려하여 훈련을 통해 식 (5)와 같이 Dirichlet density function으로 모델링 할 수 있다.

$$\rho(f_1,f_2,\cdots,f_{r-1}) \tag{5}$$
$$= Dir(f_1,f_2,\cdots,f_{r-1};a_1,a_2,\ldots,a_r)$$
$$= \frac{\Gamma(M)}{\prod_{i=1}^{r}\Gamma(a_i)} f_1^{a_1-1} f_2^{a_2-1} \cdots f_r^{a_r-1}$$

$$0 \le f_k \le 1,\ \sum_{k=1}^{r} f_k = 1_k,\ M = \sum_{k=1}^{r} a_k$$

결국 다채널 마이크 환경에서 통합 인식결과를 얻기 위해 훈련해야 할 PDF는 그림 3과 같다. 그림에서 보는바와 같이 각 node의 PDF Function은 node의 parent을 고려한 훈련을 통하여 얻을 수 있다.

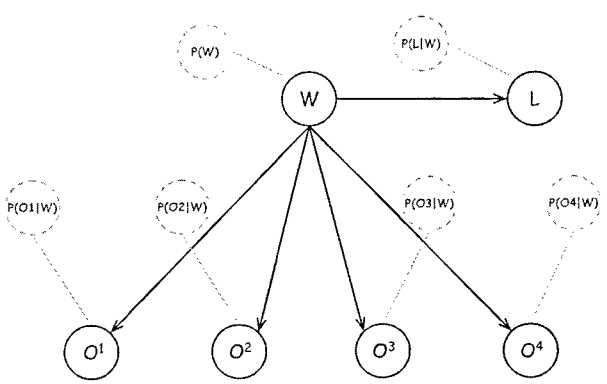

그림 4. Naive Bayesian Network을 이용한 N-best 인식결과 통합을 위해 필요한 PDF function

Ⅲ. 실험 및 결과

1. Database

각 마이크로부터의 인식결과를 통합하기 위한 integration 방법의 성능평가를 위해 그림 1과 같은 실험실 환경에서 방안을 제어하기 위한 30개의 단어를 선

별하여 녹음하고 그 성능을 평가하였다. 셀의 양 끝에 지향성 마이크를 설치하고 그 방향이 셀의 중심을 향하도록 하였고, 그림에서 보는 바와 같이 총 5곳의 위치에서 각각 5 세트씩을 녹음하였다. 남자 12명과 여자 4명 총 16명의 화자를 녹음하되 실험실 안의 컴퓨터에 의한 잡음을 포함하여 녹음하였다. 그 중 9명의 남자와 3명의 여자 분의 DB를 사용하여 훈련하고 나머지 3명의 남자와 1명의 여자 분을 성능을 평가하는데 사용하였다. context-dependent 모델을 사용하였으며 Gaussian mixture 1을 사용하여 훈련하였다. 자세한 설명은 표 1과 같다.

표 1. 훈련 및 테스트 DB

발화 수	약 48,000 단어 16(명)x30(어휘)x5(위치)x5(회)x4(마이크)
잡음상황	실험실 환경으로 여러 대의 PC가 켜있는 상황
Sampling	16 kHz
화자	총 16명(남자 12명, 여자 4명) - (훈련) 12명 (남자 9명, 여자 3명) - (테스트) 4명 (남자 3명, 여자 1명)
어휘	총 30 단어
위치	그림 1에서 표시된 5곳
마이크	총 4 벌 (지향성 마이크)

2. 실험결과

앞서 언급한 바와 같이 다채널 마이크 환경에서 인식성능을 높이기 위해 각 채널의 N-best 결과를 통합하는 방법의 성능을 표 3에서 비교하였다. Base란 마이크별 성능, 즉 4개의 마이크에 의한 각 채널별 인식률을 의미하고, ML이란 4개의 마이크 별 인식결과의 likelihood 값의 크기에 따라 통합한 결과를 나타내고, LI란 2장의 1절에서 자세히 설명된 likelihood 기반의 통합방법을 의미한다. 마지막으로 BN에서는 naive BN을 사용하여 채널의 N-best 결과를 통합하였다. 결과에서 보듯 BN, LI, ML 순으로 인식 성능이 높았다.

표 2. 각 채널의 N-best 결과 통합 알고리즘에 따른 인식 성능평가

MIC# / TYPE	CH#1	CH#2	CH#3	CH#4
Base	99.22	97.12	97.94	99.53
ML	94.65			
LI	96.18			
BN	96.21			

Ⅳ. 결론

유비쿼터스 환경에서 hands-free 응용을 현실화 할 수 있는 핵심 기술인 다채널 마이크 환경에서 인식결과를 통합하는 기술을 제안하고 그 성능을 평가하였다. N-best 결과를 통합하는 방법을 제안하고 성능을 비교하였다. 결과에서 보듯 naive BN을 사용한 통합방법이 가장 높은 성능을 보였다. 또한 단순히 maximum likelihood에 의한 결정보다는 likelihood 기반의 통합방법을 사용하는 것이 더 나은 성능을 보였다.

Baseline의 성능이 높아 통합 방법의 보다 정확한 성능 비교에 어려움이 있었다. 따라서 향후계획으로는 가정환경 내에서의 잡음을 고려하여 잡음 DB를 구축하여 추가실험을 통해 BN을 이용한 N-best 결과 통합 방법의 성능을 재확인할 것이다. 또한 다른 종류의 특징도 사용할 것이며 naive BN에서 발전된 BN 구조로 발전시켜 나갈 것이다.

참고문헌

[1] Q. Lin et al., "Experiments on distant-talking speech recognition," *Proc. Spoken Language Systems Technology Workshop*, pp. 187-192, 1995.

[2] T. B. Hughes et al., "Using a real-time, tracking microphone array as input to an HMM speech recognizer," *Proc. of ICASSP*, Vol. 1, pp. 249-252 1998.

[3] T. Yamada et al., "Hands-free speech recognition with talker localization by a micophone array," *Trans. Information Processing Society of Japan*, Vol. 39, no. 5, pp. 1275-1284, 1998.

[4] T. Yamada et al., "HANDS-FREE SPEECH RECOGNITION ON 3-d VITERBI SEARCH USING A MICROPHONE ARRAY," *Proc. IEEE ICASSP*, Vol. 1, pp. 245-248, 1998.

[5] Y. Shimizu et al., "SPEECH RECOGNITION BASED ON SPACE DIVERSITY USING DISTRIBUTED MULTI-MICROPHONE," *Proc. IEEE ICASSP*, Vol. 3, pp. 1747-1750, 2000.