모노폰 거리를 이용한 트라이폰 클러스터링 방법 연구

방 규 섭, 육 동 석*
고려대학교 컴퓨터학과

Efficient Triphone Clustering Using Monophone Distance

Kyuseop Bang, Dongsuk Yook
Department of Computer Science and Engineering, Korea University
E-mail : eoul@voice.korea.ac.kr, yook@voice.korea.ac.kr

Abstract

The purpose of state tying is to reduce the number of models and to use relatively reliable output probability distributions. There are two approaches: one is top down clustering and the other is bottom up clustering. For seen data, the performance of bottom up approach is better than that of top down approach. In this paper, we propose a new clustering technique that can enhance the undertrained triphone clustering performance. The basic idea is to tie unreliable triphones before clustering. An unreliable triphone is the one that appears in the training data too infrequently to train the model accurately. We propose to use monophone distance to preprocess these unreliable triphones. It has been shown in a pilot experiment that the proposed method reduces the error rate significantly.

I. 서론

음성인식에서 음향 모델의 단위로는 음소(phoneme), 음절(syllable), 단어(word) 등이 있다. 단어 모델은 음운 변화 현상을 잘 반영한 모델로 소어휘 단어 기반 인식에서 좋은 성능을 보인다. 하지만 대어휘 음성인식에서는 단어 모델이 학습 데이터의 단어를 공유할 수 없기 때문에 개별적으로 학습된다. 따라서 적절한 학습을 위해서는 각 단어별로 충분한 학습 데이터가 존재해야한다. 음절은 단어보다 작은 단위의 모델로 음운 변화 현상을 잘 표현할 수 있다. 하지만 음절의 앞, 뒤 부분에 미치는 음운 변화 현상을 잘 표현 못할 뿐만 아니라, 모델의 개수가 많아서 학습 데이터를 공유하기 힘들다. 반면, 음소는 전체 모델의 개수가 적기 때문에 적은 학습 데이터로 충분히 학습시킬 수 있다. 하지만 음소는 주변 문맥에 의한 음운 변화 현상을 제대로 반영하기 힘들다. 따라서 대어휘 연속 음성인식에서는 주로 좌우 문맥을 고려한 트라이폰을 사용한다. 하지만 트라이폰은 모델의 개수가 많아서 메모리 소비가 심하며 정확히 학습시키기 위해서 많은 학습 데이터가 필요하다[1]. 그래서 트라이폰을 인식 단위로 사용하기 위해서는 트라이폰(triphone) 클러스터링이 필요하다[2][3][4][5]. 트라이폰 클러스터링은 음향학적인 특성이 비슷한 것을 하나로 묶거나 음성학적 특성이 비슷한 모델들을 공유함으로써 데이터 부족 현상을 해결할 수 있는데, 트라이폰 클러스터링 알고리즘은 하향식 알고리즘과 상향식 알고리즘이 있다.

상향식 알고리즘은 각 트라이폰을 개별적으로 두고 트라이폰들간의 거리를 비교하고 가장 비슷한 트라이폰을 묶는 방법이다. 이렇게 묶인 트라이폰은 더 많은 학습 데이터로 정확하게 학습시킬 수 있다[2][3]. 반면 하향식 알고리즘은 언어학적 정보를 통해 비슷한 문맥을

* 교신저자
본 연구는 한국과학재단 특정기초연구 (R01-2006- 000-11162-0) 지원으로 수행되었음.

미리 정한 뒤 결정 트리의 각 노드의 질문으로 사용하여 클러스터링 하는 방법이다[4][5]. 일반적으로 상향식 알고리즘이 보다 강인하게 묶을 수 있지만 하향식 알고리즘이 더 좋은 성능을 보인다. 이는 unseen 트라이폰과 신뢰할 수 없는 트라이폰때문 인데, unseen 트라이폰은 학습 데이터에서는 나타나지는 않지만 인식과정에 나타나는 것이고, 신뢰할 수 없는 트라이폰은 학습데이터에 나타나긴 하지만 출현 빈도수가 적어서 정확하게 모델링할 수 없는 것을 말한다. 상향식 알고리즘은 이들을 모노폰이나 다이폰으로 대체하여 사용하지만[2][3], 하향식 알고리즘은 결정 트리를 통해 비슷한 트라이폰으로 대체한다[4][5]. 만약 위에서 언급한 unseen 트라이폰과 신뢰할 수 없는 트라이폰 문제를 해결 한다면 상향식 알고리즘이 하향식 알고리즘보다 더 좋은 성능을 보일 것이다. 본 논문에서는 위에서 언급한 신뢰할 수 없는 트라이폰을 효율적으로 다룰 수 있는 상향식 알고리즘을 제시한다.

본 논문은 2장에서 기존 클러스터링 알고리즘을 소개하고 3장에서 신뢰할 수 없는 트라이폰을 처리하는 preprocessing에 대해 설명한다. 그리고 4장에서는 실험환경 및 결과를 제시하고 5장에서는 결론을 맺는다.

II. 하향식 알고리즘 vs. 상향식 알고리즘

2.1 상향식 알고리즘

상향식 알고리즘은 흔히 data-driven 방식이라고 하기도 한다. 상향식 알고리즘은 2단계로 나눌 수 있다[6].

1) 가장 유사한 즉, 거리값이 작은 두개의 가우시안(Gausian)을 묶는다. 이 과정은 거리값이 일정 threshold를 만족할 때까지 수행한다.
2) 출현 빈도수를 나타내는 γ가 작은 가우시안을 가장 유사한 가우시안과 묶는다. 이 과정은 모든 클러스터의 γ가 일정 threshold를 만족할 때까지 수행한다.

상향식 알고리즘은 학습 데이터로부터 계산되어지기 때문에 신빙성이 높은 장점이 있다. 따라서 두 클러스터가 언어학적 규칙상 비슷하다고 해서 두 클러스터간의 거리가 가깝지 않으면 클러스터링 되지 않을 수도 있다[7]. 그러나 출현빈도수가 낮아서 정확하게 학습되지 않은 트라이폰이나 학습 데이터에 나타나지 않은 unseen 트라이폰 효과적으로 처리할 수 없다. 그래서 바이폰이나 모노폰으로 대체하는데 이로 인해 인식 성능의 저하가 생긴다[4][5].

2.2 하향식 알고리즘

하향식 알고리즘은 흔히 결정 트리기반 혹은 규칙기반 알고리즘이라고 부르기도 한다. 하향식 알고리즘은 다음과 같다[4][5].

1) 언어학적 질문에 따라 각 음소들을 구분한다.
2) 모든 트라이폰을 결정트리의 루트(root)에 놓고 다음과 같이 분기한다.
   - 각 노드들중 분기하기에 가장 적합한 질문을 찾고 그 질문에 따라 노드를 자식노드로 분기한다.
   - 정지 조건을 만족할 때까지 위의 과정을 반복한다.
3) 단말 노드의 가우시안들을 하나의 클러스터로 구성한다.

하향식 알고리즘은 학습되지 않은 unseen 트라이폰이나 신뢰할 수 없는 트라이폰을 결정 트리상의 다른 트라이폰으로 대체함으로써 효과적으로 unseen 트라이폰과 신뢰할 수 없는 트라이폰 문제를 해결할 수 있다. 하지만 언어학적 질문에 기반을 둔 결정트리를 사용하기 때문에 상향식 알고리즘보다 덜 강인하게 클러스터링 된다. Seen 트라이폰에 대해서는 상향식 알고리즘이 하향식 알고리즘보다 우수한 성능을 나타낸다[5]. 따라서 본 논문에서는 트라이폰을 상향식 알고리즘보다 더욱 더 강인하게 묶을 수 있는 방법을 제시한다.

III. 제안한 알고리즘에 의한 상향식 클러스터링 알고리즘

트라이폰 클러스터링은 정확한 음향 모델을 만들기 위해 반드시 필요한 알고리즘이다. 그러나 신뢰할 수 없는 트라이폰이나 unseen 트리이폰으로 인해 음성인식 성능 저하가 일어난다. 따라서 이를 처리할 수 있는 효과적인 알고리즘이 필요하다. 하향식 알고리즘은 결정 트리에서 감마값이 threshold보다 작으면 분기하지 않는 방법으로 신뢰할 수 없는 트라이폰을 처리할 수 있다. 하지만 상향식 알고리즘은 각 트라이폰들간의 거리값을 이용하여 비슷한 트라이폰을 묶어나가기 때문에 각 트리아폰을 신뢰할 수 없다면 효과적인 클러스터링을 할 수 없다.

3.1 Unreliable Triphone Preprocessing

Unreliable triphone preprocessing(UTP)은 하향식 알고리즘을 수행하기 전에 신뢰할 수 없는 데이터들을 트라이폰의 거리값이 아닌 다른 정보를 이용하여 먼저 묶는 방법이다. 트라이폰의 좌, 우 문맥은 각각 기본 음소에 음운 변화를 일으킨다. 이러한 기본 음소에 미치는 음운 변화 영향을 모노폰의 거리를 통해 표현할 수 있다. 모노폰의 거리를 이용하여 두 클러스터 C_1, C_2의 거리를 구하는 공식은 식 (1)과 같다.

$$d(C_1, C_2) = \frac{\sum_{m \in C_1} \sum_{n \in C_2} d_{tri}(m,n)}{\sum_{m \in C_1} \sum_{n \in C_2}} \quad (1)$$

$$d_{tri}(m,n) = w_l d_m(m^l, n^l) + w_r d_m(m^r, n^r) \quad (2)$$

식 (1)에서 C_1, C_2는 각각 클러스터를, m, n은 클러스터에 속해있는 트라이폰를 나타낸다. 그리고 식 (2)는 두 트라이폰의 거리값을 모노폰의 거리값을 이용하여 나타낸 식이다. m^l과 m^r은 각각 트라이폰의 좌, 우 문맥을 의미하고, w_l과 w_r는 각각의 가중치 값이다. 트라이폰의 거리값은 각 트라이폰의 좌, 우 문맥에 대한 모노폰 거리값에 가중치를 곱하여 더한 값으로 표현된다. 여기서 사용되는 모노폰 거리값은 하나의 state에 하나의 가우시안이 존재한다고 가정하고 식 (3)의 Bhattacharyya 거리값을 이용하여 구하였다.

$$B(g,h) = \frac{1}{8}(\mu_g - \mu_h)^T \left|\frac{\Sigma_g + \Sigma_h}{2}\right|^{-1} (\mu_g - \mu_h) + \frac{1}{2} \frac{\left|\frac{\Sigma_g + \Sigma_h}{2}\right|}{\sqrt{|\Sigma_g||\Sigma_h|}} \quad (3)$$

식 (3)에서의 μ와 Σ는 가우시안 g와 h의 평균 벡터와 공분산 행렬을 뜻한다.

Unreliable 트라이폰은 위의 식을 이용하여 먼저 처리할 수 있는데, UTP 알고리즘은 다음과 같다.

1. 초기화 단계
   - 각 클러스터에 하나의 트라이폰이 놓이도록 구성하고, 전체 클러스터 집합 C와 unreliable 트라이폰이 속한 C_u로 구분한다.
2. Preprocessing 단계
   2-1. Merge 단계
   - C_u의 클러스터와 C의 클러스터의 모든 쌍의 거리를 식 (1)을 이용하여 구한다.
   - 위의 쌍 중에서 가장 거리값이 작은 두 클러스터를 묶어 새로운 클러스터 C_n을 만든다.

   2-2. Reclassification 단계
   - 위에서 묶여진 두 클러스터를 C와 C_u에서 삭제한다.
   - C_n의 γ가 threshold보다 낮으면 C_n을 C_u에 추가하고 threshold보다 높으면 C에 추가한다.
3. 종료 조건
   - C_u의 원소 개수가 0이 아니면 2단계를 반복 수행한다.

Ⅳ. 인식실험 및 결과

제안된 UTP기반 상향식 알고리즘의 성능을 평가하기 위해 TIMIT 데이터를 사용하였다. 그리고 unseen 트라이폰이 없는 환경을 만들기 위해 실험 데이터는 TIMIT 학습 데이터 중 dr8을 테스트 데이터로 사용하여 unseen 트라이폰이 없는 환경을 만들었다. 음성 특징 벡터로는 매 10ms마다 멜 스케일 켑스트럼 계수에 에너지 값을 추가한 13차원 벡터와 그 1,2차 미분값이 더해진 총 39차원의 벡터를 추출하여 사용하였다. HMM의 각 상태는 10개의 mixture들로 구성된다.

그림 4-1은 모노폰들간의 거리를 Bhattacharyya 거리값으로 표현한 것이다.

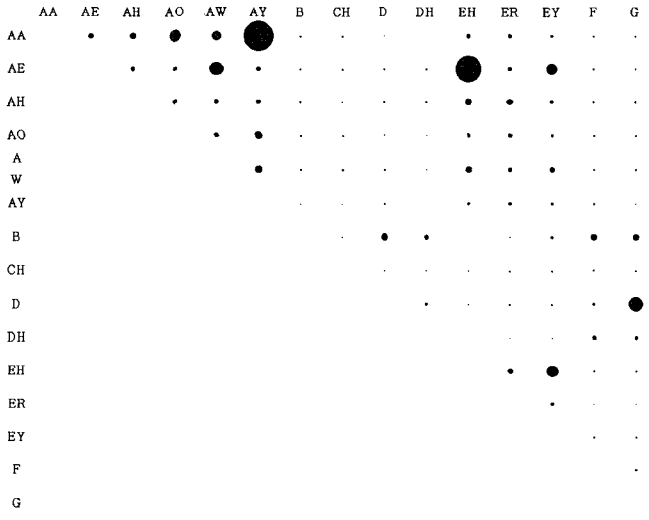

그림 4-1. Bhttacharyya 거리값으로 나타낸 모노폰들 간의 거리

두 모노폰의 거리가 가까울수록 검정색 원이 커진다. 두 모노폰 사이의 거리는 음향적 특성과 유사형태를 보였다. 특히 높은 친밀도를 보인 AA-AY, AE-EH, D-G 등은 하향식 알고리즘에서 사용하는 음향적 특성에 의한 분류과 같은 형태였다. 이것은 Bhattacharyya 거리

값이 모노폰들간의 친밀도를 표현하는데 좋은 척도임을 나타낸다.

하향식 알고리즘은 로그 우도확률 기반으로 결정트리를 구축하여 6.8% 에러율을 보였다. 이와 같이 학습된 모델을 이용하여 인식 실험 결과 표 4-1에서 보듯이 기존 하향식 알고리즘에 비해 31.9%의 에러율이 감소되었으며, 기존 상향식 알고리즘과 비교하여 25.4%의 에러율 감소를 보여 본 논문에서 제안한 UTP기반 상향식 알고리즘이 강인한 트라이폰 클러스터링을 할 수 있음을 보였다.

클러스터링 알고리즘	Error rate (%)
하향식 알고리즘	6.9
상향식 알고리즘	6.3
UTP 기반 상향식 알고리즘	4.7

표 4-1. 클러스터링 알고리즘 별 에러율

D-F+AH의 경우 하향식 알고리즘과 상향식 알고리즘에서는 왼쪽 문맥이 CH, D, DH, JH SH, ZH인 트라이폰과 같은 클러스터를 이루었다. 하지만 UTP기반 상향식 알고리즘은 CH, JH, SH, ZH를 한 클러스터로, D와 DH를 다른 클러스터로 구성하여 D에 의해 일어나는 음운 변동을 더 구체적으로 표현할 수 있다. 이로 인해 기존 알고리즘에서 발생하였던 에러가 발생하지 않았다.

	하향식 알고리즘	상향식 알고리즘	UTP 기반 상향식 알고리즘
AA-F+AH			o
AO-F+AO	o	o	o
AY-F+AH	o		o
D-F+AH	o	o	
ER-F+AH	o		
IH-F+R	o	o	o
IH-F+S	o	o	
IH-F+Y	o	o	
M-F+AH	o	o	
N-F+AO	o		o
P-F+IH	o	o	o
R-F+AH	o		
R-F+AO		o	
T-F+ER	o		
T-F+R	o		
Z-F+ER		o	
Z-F+IH		o	
총 계	13	10	6

표 4-2. 각 알고리즘별 에러를 일으키는 트라이폰

표 4-2는 기본 음소 F에 대한 각 알고리즘별 에러를 일으키는 트라이폰을 표시한 표이다. 위에서 설명한 것과 같이 UTP기반 상향식 알고리즘에서 에러를 일으키는 트라이폰의 수가 기존 알고리즘에 비해 줄었음을 알 수 있다.

V. 결론

본 논문에서는 신뢰할 수 없는 트라이폰을 모노폰 거리를 이용하여 처리함으로써 기존의 상향식 알고리즘보다 트라이폰을 강인하게 묶을 수 있는 알고리즘을 제시하고 인식 실험을 통해 기존의 알고리즘과 비교 분석하였다. Unseen 트라이폰이 없는 환경에서는 인식 성능이 향상되어 강인한 클러스터링이 이루어졌음을 증명하였다. 숫자음 인식이나 인식 어휘가 한정되어 있어 unseen 트라이폰이 거의 없는 경우에는 유용한 알고리즘이다.

참고문헌

[1] K. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 599-609, 1990.

[2] M. Hwang, and X. Huang, "Shared-distribution hidden Markov models for speech recognition", *IEEE Transaction on Speech and Audio Processing*, vol. 1, No. 4, 1993.

[3] S. Young and P. Woodland "State clustering in hidden Markov model-based continuous speech recognition", *Computer Speech and Language*, vol. 8, pp. 369-383, 1994.

[4] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling", *DARPA Human Language Technology Workshop*, pp. 307-312, March 1994.

[5] M. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones", *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 412-419, November 1996.

[6] J. Odell, "The use of context in large vocabulary speech recognition", PhD thesis, University of Cambridge, 1996.

[7] J. Park and H. Ko, "Construction of decision tree form data driven clustering", ICSLP 2002, pp. 2657-2660, 2002.