분산 음성인식 시스템의 성능향상을 위한 음소 빈도 비율에 기반한 VQ 코드북 설계

오유리[1], 윤재삼[1], 이길호[2], 김홍국[1], 류창선[3], 구명완[3]
[1]광주과학기술원 정보통신공학과, [2]삼성전자 CS 경영센터, [3]KT 음성언어연구부

A VQ Codebook Design Based on Phonetic Distribution for Distributed Speech Recognition

Yoo Rhee Oh[1], Jae Sam Yoon[1], Gil Ho Lee[2], Hong Kook Kim[1], Chang-Sun Ryu[3], and Myoung-Wan Koo[3]
[1]Dept. of Information and Communications, Gwangju Institute of Science and Technology
[2]CS Management Center, Samsung Electronics
[3]Speech Research Division, Spoken Language Research Department, Advanced Technology Laboratory Kora Telecom
E-mail : {[1]yroh, [1]jsyoon, [1]hongkook}@gist.ac.kr,
[2]ghlee@gmail.com, {[3]csryu, [3]mwkoo}@kt.co.kr

Abstract

In this paper, we propose a VQ codebook design of speech recognition feature parameters in order to improve the performance of a distributed speech recognition system. For the context-dependent HMMs, a VQ codebook should be correlated with phonetic distributions in the training data for HMMs. Thus, we focus on a selection method of training data based on phonetic distribution instead of using all the training data for an efficient VQ codebook design. From the speech recognition experiments using the Aurora 4 database, the distributed speech recognition system employing a VQ codebook designed by the proposed method reduced the word error rate (WER) by 10% when compared with that using a VQ codebook trained with the whole training data.

I. 서론

휴대폰, PDA 등 휴대 전자기기의 사용 증가로 인하여 무선 네트워크 환경에서 음성인식 서비스를 효과적으로 제공하기 위한 분산 음성인식시스템에 대한 필요성이 날로 증대되고 있다. 분산 음성인식시스템은 단말기 측에서 음성의 특징벡터를 추출하고 압축과정을 거쳐 서버 측으로 음성의 특징벡터를 전송하고, 서버 측에서는 전송받은 데이터를 이용하여 음성인식을 수행하는 시스템이다.

분산 음성인식시스템의 단말기 측에서 음성의 특징벡터를 전송하기 전에 코드북을 이용하여 압축과정이 수행된다. 여기서 문맥종속형 HMM (hidden Markov model)에 기반한 음성인식의 경우, VQ (vector quantizer)의 코드북은 HMM에서 음소의 표현과 밀접한 관계가 있어야 한다. 따라서 본 논문에서는 무작위로 수집된 학습용 데이터 전체를 사용하여 VQ 코드북을 생성하는 대신, 학습용 데이터에서의 음소 출현 빈도를 고려하여 학습용 데이터를 선별하여 VQ 코드북을 효율적으로 생성하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 분산 음성인식시스템에서 사용된 VQ를 소개

하고, 음소 출현 빈도 분석을 통하여 선정된 학습용 데이터를 이용하여 VQ 코드북을 생성하는 방법을 제안한다. 그리고 3장에서는 제안한 방법에 대한 분산 음성인식 실험 및 그 결과를 보여주고, 4장에서 결론을 맺는다.

II. VQ 코드북 설계

본 장에서는 먼저 분산 음성인식시스템에서 사용하는 특징벡터 전송을 위한 유럽 표준 front-end 알고리즘을 설명한 후, 분산 음성인식시스템의 음소 빈도 비율 분석에 기반한 VQ 코드북 설계 방법을 제안한다. 제안하는 VQ 코드북 설계는 VQ 코드북의 학습 데이터 선정 방법에 초점을 둔다.

2.1. VQ 구조

본 장에서는 단말기 측과 서버 측으로 구성되는 분산 음성인식시스템의 표준 front-end 알고리즘인, ETSI ES 202 050 v.1.1.3 [1]에 대하여 설명한다. 이 표준 front-end 알고리즘에서 입력 음성은 8 kHz, 11 kHz, 16 kHz 중 하나로 표본화된다. 그리고 표본화된 음성 파형으로부터 13개의 정적 켑스트럼 계수와 로그에너지를 추출한다.

그림 1은 분산 음성인식시스템의 표준 front-end 알고리즘의 구성도를 나타내며, 그림 1 (a)와 (b)는 각각 분산 음성인식시스템의 단말기 측과 서버 측에서의 음성특징 추출/압축과정과 음성특징 복원과정의 흐름도를 각각 보여준다. 그림 1 (a)에서와 같이 단말기 측에서는, 서버 측으로 음성특징을 전송하기 위한 처리를 수행한다. 먼저 특징추출 (feature extraction) 모듈에서는

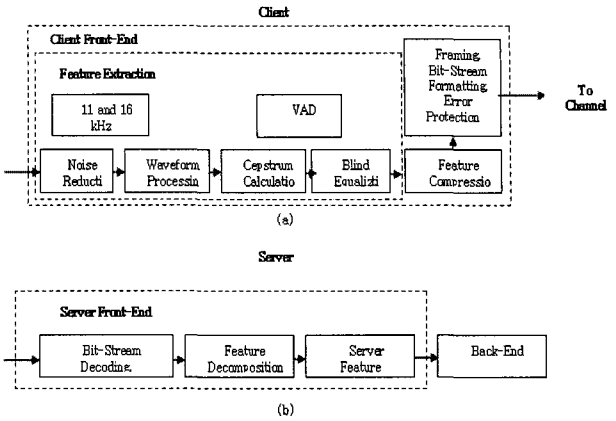

그림 1. 분산 음성인식시스템에서 특징벡터 전송을 위한 ETSI 표준 front-end 알고리즘 흐름도: (a) 단말기 측, (b) 서버 측.

입력 음성신호의 잡음을 제거한다. 그 후, 잡음이 제거된 음성 신호에 waveform processing을 적용하고 13개의 정적 켑스트럼 계수와 로그에너지를 추출한다. 마지막으로 blind equalization이 켑스트럼 특징들에 적용된다. 그림 1 (a)에서 볼 수 있듯이 특징추출 모듈은 11 kHz와 16 kHz의 확장 블록을 포함한다. 추출된 음성특징들은 feature compression 모듈을 통하여 압축되고, error protection을 할 수 있도록 비트스트림을 생성하여 서버 측으로 전송한다. 반면, 그림 1 (b)와 같이 서버 측에서는, 단말기 측으로부터 전송받은 비트스트림을 복호화하면서 에러처리도 병행한다. 그 후 전송받은 특징벡터의 압축 풀기가 수행된다. 또한 back-end (음성인식기)에 들어가기 전에, 서버 측에서 추가적인 처리가 이루어진다.

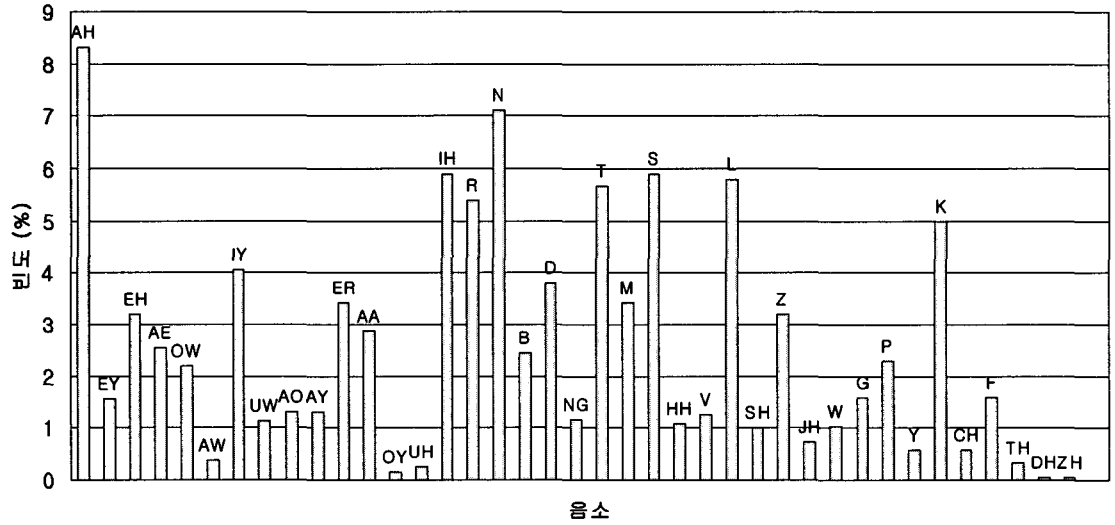

그림 2. CMU dictionary에 기반한 영어 음소 빈도, 여기서 각 음소는 2자리 ARPAbet으로 표현된다.

2.2. 음소 빈도 비율에 기반한 VQ 코드북 설계

VQ 코드북을 설계하기에 앞서, 영어 음소 빈도 비율을 조사하기 위하여 CMU dictionary v.0.6 [2]를 사용하였다. CMU dictionary는 Carnegie Mellon 대학교에서 제공하는 129,482개의 단어에 대한 발음 사전이다. CMU dictionary는 25개의 자음, 14개의 모음의 총 39개의 음소로 구성되며, 강세를 0(강세 없음), 1(주 강세), 2(보조 강세)로 구분하여 표기한다. 본 논문에서는 강세에 대한 정보는 사용하지 않았다. 그림 2는 CMU dictionary를 분석하여 얻은 영어 음소 빈도 비율을 나타내고, 여기서 각 음소는 2 자리 ARPAbet으로 표현되었다. 그림 2로부터 영어의 경우 자음과 모음의 분포는 각각 약 56%와 44%이며, 슈아(schwa)에 해당하는 AH가 8.3%로 가장 높은 빈도를 보임을 알 수 있다. 또한 AW, OY, UH, JH, Y, TH, DH, ZH 등은 CMU dictionary에서 발생 빈도가 1%를 넘지 않음을 알 수 있다.

본 논문에서 제안하는 VQ 코드북 설계 방법은 분산 음성인식시스템에서 사용되는 음소의 빈도 비율에 기반하여 VQ 코드북의 학습용 데이터를 선정하는 방법이다. 그림 3은 제안하는 음소 빈도 비율에 기반하여 VQ 코드북을 설계하는 과정에 대한 흐름도이다. 먼저, 각 음성에 대한 음소 정보를 얻기 위하여 음성인식시스템을 생성한다. 그 후, 무작위로 수집된 학습용 데이터 전체의 각 음성 파일에 대한 음소 정보 (음소별 시작시간, 종료시간 등)를 얻기 위하여 forced alignment를 수행하고, 각 음소에 대하여 시작시간 종료시간 등을 리스트로 생성한다. 마지막으로 학습용 데이터의 각 음소에 대한 전체 빈도 비율과, 통계적 분석을 통하여 얻은 일반적인 음소 빈도 비율을 이용하여 오차를 최소한으로 하도록 학습 데이터를 선정한다. 단, 본 논문에서는 음소별 출현 빈도에만 초점을 두었으며, 음소의 길이 및 음소의 특성 등에 대한 것은 고려되지 않았다. 예를 들어, 어떤 음소 /a/에 대한 영어 빈도 비율이 2%이고, 음성인식시스템의 학습용 데이터에서 음소 /a/가 발생

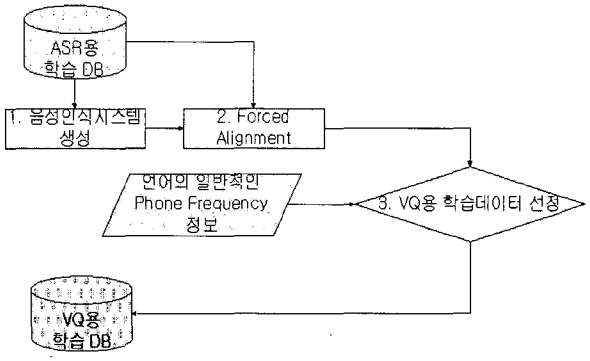

그림 3. VQ 코드북의 학습 데이터 선정 흐름도.

한 횟수가 40이며 학습용 데이터에서 발생한 총 음소 수가 1,000개라고 가정하자. 이 경우 학습용 데이터에서 음소 /a/가 발생한 빈도 비율이 4% 이다. 그러므로 forced alignment 후 생성한 음소 /a/에 대한 리스트 중 1, 3, 5,…, 39 번째 데이터만을 선정하여 VQ 코드북의 학습용 데이터로 사용한다.

III. 실험

3.1. 영어 음성 데이터베이스

Wall Street Journal 데이터베이스 [3] (WSJ0) 의 multi-condition 학습용 데이터가 영어 음성인식시스템의 학습용 데이터로 사용하였다. WSJ0은 대어휘 연속 음성인식의 성능을 평가하기 위하여 구축되었으며 5,000단어로 구성되어 있다. 영어음성인식시스템의 학습용 데이터는 Sennheiser 근거리 마이크와 몇 가지 종류의 원거리 마이크로 녹음된 7,138개의 발화음성으로 구성되며, 16 kHz의 샘플링레이트로 저장되었다.

또한 음성인식시스템의 성능 평가를 위하여 5,000 단어 급에 해당하는 WSJ0의 평가용 데이터 중에서 multi-condition 평가용 데이터 14개의 Set 중 일부인 Set1 ~ Set7을 사용하였다. 각 Set은 330개 음성 데이터로 구성되며, 이는 8명의 화자로부터 녹음된 발화 음성으로 한 화자 당 약 40개의 음성을 발화하였다.

3.2. 영어 음성인식시스템

영어 음성인식시스템에는 39차 특징 벡터가 사용되었으며, 이를 위하여 12차 멜-켑스트럼 계수 (MFCC), 로그 에너지를 추출하였고, 1차, 2차 미분계수를 사용하였다. 특징 벡터 추출을 위하여 2.1 장에서 설명된 ETSI 분산음성인식 front-end 표준 알고리즘을 이용하였다. 또한 인식 및 학습에 사용된 특징벡터에 켑스트럼 평균 정규화와 에너지 정규화 기법을 적용하였다.

음향 모델은 3개의 state를 갖는 left-to-right model이 사용되었으며, 문맥 독립적이고, 4개의 혼합밀도와 cross-word 트라이폰 모델을 사용하였다. 또한 음향 모델은 HTK version 3.2 toolkit [4]을 이용하여 학습되었다. 먼저 두 개의 묵음 모델이 포함된 41 개의 단음에 기반을 둔 음향모델에서 시작하여 트라이폰에 기반을 둔 음향모델로 확장한 후, 이진트리를 이용한 상태 공유 단계를 거쳐 상태의 수를 줄였다. 그 결과 영어 음성인식시스템은 8,360개의 트라이폰과 5,356 상태의 음향 모델로 구성되었다.

우선 baseline으로, 벡터양자화는 수행되지 않은 인식시스템의 성능을 평가하였다. 표 1은 본 장에서 소개한 baseline 영어 음성인식시스템의 단어오인식률 (%)을

표 1. VQ를 사용하지 않은 경우, 영어 음성인식시스템에서 각 평가용 세트 단어오인식률 (%)과 평균 단어오인식률 (%)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Avg
DSR 벡터 양자화기를 사용하지 않은 음성인식시스템	14.3	17.6	22.1	27.0	25.1	23.1	23.5	21.8

나타내는 것으로, 7개의 Set에 대하여 평균 단어오인식률이 21.8%임을 보인다.

3.2 제안한 VQ 코드북 설계 방법에 대한 분산 음성인식 실험 결과

표 2. WSJ0의 multi-condition 학습데이터 전체로 부터 작성된 VQ 코드북을 사용한 분산 음성인식시스템의 단어오인식률 (%)과 제안한 방식으로 학습용 데이터를 선정 후 작성된 VQ 코드북을 사용한 분산 음성인식시스템의 단어오인식률 (%)

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Avg
학습데이터 전체를 VQ 코드북 생성에 사용한 경우	17.8	18.9	25.1	29.4	26.5	24.8	25.9	24.1
전체 학습데이터 중에서 음소 빈도 비율에 기반하여 트레이닝 데이터를 선정한 경우	13.6	16.8	23.0	27.6	24.9	21.8	24.5	21.7

먼저, 제안하는 VQ 코드북의 성능을 비교하기 위하여 WSJ0의 학습데이터 중 multi-condition 학습용 데이터 전체를 이용하여 VQ 코드북을 생성하였다. 그리고, WSJ0의 multi-condition 학습용 데이터 중에서 그림 2의 영어 음소 빈도 비율을 바탕으로 VQ 코드북 학습용 데이터를 선정한 후 VQ 코드북을 생성하였다.

표 2는 WSJ0의 학습데이터 중 multi-condition 학습데이터를 모두 사용하여 VQ 코드북을 생성하였을 때와, 제안한 VQ 코드북 설계 방법에 의하여 VQ 코드북을 생성하였을 때의 평가용 세트 Set1~Set7에 대한 단어오인식률 및 평균 단어오인식률을 나타낸다. WSJ0의 학습데이터 중 multi-condition 학습데이터를 모두 사용하여 VQ 코드북을 생성한 경우 7개의 Set에 대한 평균 단어오인식률은 24.1%이다. 반면, 제안한 음소 빈도 비율에 기반한 VQ 코드북으로 VQ 코드북을 생성한 경우 7개의 Set에 대한 평균 단어오인식률은 21.7%이다. 다시 말해, 제안한 VQ 코드북 방법으로 학습데이터를 선정하여 VQ 코드북을 생성한 경우, 전체 학습데이터를 사용한 경우와 비교하였을 때 단어오인식률이 24.1%에서 21.7%로 향상되었고, 단어오인식률이 약 10% 감소하였음을 알 수 있다. 뿐만 아니라, 제안한 VQ 코드북 설계방법으로 분산 음성인식시스템을 생성하였을 때, 특징벡터를 양자화하지 않은 경우의 단어오인식률 21.8%와 유사함을 알 수 있다.

그러므로 분산 음성인식 실험 결과로부터 본 논문에서 제안한 음소 빈도 비율에 기반한 VQ 코드북 설계 방법은 분산 음성인식시스템의 인식성능을 향상시킬 수 있음을 알 수 있다.

IV. 결론

본 논문에서는 분산 음성인식시스템의 성능향상을 위하여 음소 빈도 비율에 기반하여 음성인식 특징파라미터의 벡터 양자화기 (VQ) 코드북 설계 방법을 제안하였다. 또한 본 논문에서 제안한 VQ 코드북 설계 방법은, 일반적으로 널리 사용하는 문맥종속형 HMM을 기반으로 하는 분산 음성인식시스템의 성능향상을 목적으로 한다. Aurora 4 데이터베이스를 가지고 분산 음성인식 실험결과, 학습용 데이터 전체를 학습하여 VQ 코드북을 설계하는 경우에 비교하여, 본 논문에서 제안한 음소 빈도 비율에 기반한 VQ 코드북 제작 방법으로 코드북을 생성하는 경우, 단어오인식률이 약 10% 정도 감소하였음을 알 수 있었다. 뿐만 아니라, 제안한 VQ 코드북 설계방법을 통하여 특징벡터를 양자화하지 않은 경우와 유사한 단어오인식률을 얻을 수 있었다.

감사의글

이 논문은 주식회사 KT의 지원과 2005-2006년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2005-202-D00367).

참고문헌

[1] ETSI ES 202 050 v.1.1.3, *Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced frond-end feature extraction algorithm; Compression algorithms*, Nov. 2003.

[2] H. Weide, "The CMU Pronunciation Dictionary, release 0.6," Carnegie Mellon University, 1998.

[3] D. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in Proc. DARPA Speech and Language Workshop, Arden House, NY, pp. 357-362, Feb. 1992.

[4] S. Young, et al, *The HTK Book (for HTK Version 3.2)*, Microsoft Corporation, Cambridge University Engineering Department, Dec. 2002.