연속음성인식의 음향모델 출력을 이용한 뉴스 데이터 분석

News Data Analysis Using Acoustic Model Output of Continuous Speech Recognition

이경록
남부대학교 디지털정보학과
Kyong-Rok Lee(krlee@nambu.ac.kr)

요약

본 논문에서는 연속음성인식의 음향모델 출력을 이용하여 뉴스 데이터를 분석하였다.

실험에 사용된 뉴스 데이터베이스는 2,093개의 기사로 구성되어 있다. 기존의 한국어 연속음성인식은 열악한 언어모델 때문에 낮은 인식성능을 보여 뉴스 데이터 분석에 적합하지 않다. 본 논문에서는 이를 보완하기 위해서 상대적으로 견인한 음향모델의 인식결과를 후처리하여 핵심어 정보 파일을 만들었다.

음향모델의 출력레벨 문턱치가 100일 때 전체 인식대상 형태소의 86.9%가 인식되었다. 동일한 조건에 길이정보 기반 정규화를 적용하였더니 81.25%가 인식되었다. 정규화의 목적은 긴 길이의 형태소를 보상하는 것이다. 실험결과, 인식대상 형태소 인식률은 75.13%였다. 그리고 5,040MB의 뉴스 데이터에서 314MB의 핵심어 정보 파일이 만들어졌다. 이것은 절대적인 정보량이 93.8% 감소한 것이다.

■ 중심어 : | 뉴스 데이터 분석 | 형태소 기반 연속음성인식 |

Abstract

In this paper, the acoustic model output of CSR(Continuous Speech Recognition) was used to analyze news data. News database used in this experiment was consisted of 2,093 articles.

Due to the low efficiency of language model, conventional Korean CSR is not appropriate to the analysis of news data. This problem could be handled successfully by introducing post-processing work of recognition result of acoustic model. The acoustic model more robust than language model in Korean environment. The result of post-processing work was made into KIF(Keyword information file).

When threshold of acoustic model's output level was 100, 86.9% of whole target morpheme was included in post-processing result. At the same condition, applying length information based normalization, 81.25% of whole target morpheme was recognized. The purpose of normalization was to compensate long-length morpheme.

According to experiment result, 75.13% of whole target morpheme was recognized. KIF(314MB) had been produced from original news data(5,040MB). The decrease rate of absolute information amount was approximately 93.8%.

■ keyword : | News Data Analysis | Morpheme Based CSR |

접수번호 : #060824-002
접수일자 : 2006년 08월 24일
심사완료일 : 2006년 10월 12일
교신저자 : 이경록, e-mail : krlee@nambu.ac.kr

I. 서 론

정보의 디지털화는 멀티미디어 데이터의 생산을 촉진시켰고, 이는 멀티미디어 데이터베이스라는 새로운 패러다임을 창조하였다. 이런 멀티미디어 데이터는 기존의 텍스트 데이터와는 달리 분류과정의 자동화가 매우 어렵다. 이에 여러 가지 방법을 이용한 멀티미디어 데이터 분석이 제안되고 있다[1][2].

본 논문에서는 멀티미디어 데이터 중에서 뉴스 데이터에 주목하였다. 공중파 방송 뉴스를 이용하여 총 51회 분량의 뉴스 데이터를 수집하여 데이터베이스화 하였다. 뉴스 데이터는 다양한 음향품질, 채널 혹은 배경음향의 상태, 숙련된 발화자의 빠른 발성 등으로 인해 멀티미디어 데이터 중에서도 고난이도의 작업을 요구한다. 뉴스 데이터 분석도구로는 한국어의 특성을 고려한 형태소 기반 연속음성인식을 이용하였다[3][4].

연속음성인식의 결과는 핵심어 정보 파일로 정리되어 차후 뉴스 데이터 검색에 사용된다.

II. 뉴스 데이터베이스

1. 뉴스 데이터 선정

멀티미디어 데이터 중에서 뉴스 데이터를 선택한 이유는 다음과 같다. 첫째, 특정 화자(앵커 등)의 발성 후에 그에 해당하는 멀티미디어 정보가 이어진다. 둘째, 트랜스크립션 데이터가 제공되어 훈련이나 평가에 도움이 된다. 셋째, 공중파 방송 3사에서 데이터를 매일 생산한다. 넷째, 멀티미디어 데이터 중에서 수요가 높다. 다섯째, 정형적이어서 규칙을 적용하기 쉽다.

뉴스 데이터는 방송 3사(MBC, KBS, SBS)의 아침과 저녁 뉴스를 대상으로 선정하였다. [표 1]에서 보는 바와 같이 17회씩으로 균등하게 배분하였다. 트랜스크립션 데이터의 유무에 따라 저녁, 아침 방송의 데이터양에 변화가 있다.

표 1. 뉴스 데이터 구성

구분	MBC (회)	KBS (회)	SBS (회)
아침 / 저녁	1 / 16	9 / 8	7 / 10

표 2. 뉴스 데이터베이스 세부사양

구분	총 뉴스	총 기사	총 문장	총 화자
수량	51	2,093	21,059	2,169

표 3. 뉴스 데이터베이스의 화자 현황

구분	앵커		리포터		시민	
	남성	여성	남성	여성	남성	여성
수	31	20	421	59	1268	370

2. 뉴스 데이터베이스 구축

뉴스 데이터는 뉴스 방송을 녹화한 다음 음향출력만을 전용녹음기로 녹음하였다. 녹음된 음향출력은 일반적인 음성인식 옵션 (8 khz, 16 bit)을 적용하여 A/D (Analog/Digital) 변환하였다.

뉴스 데이터는 화자 정보(성별, 역할)와 배경소음 정보에 따라서 수동으로 분류되었다. 이는 화자정보 및 배경소음 정보에 따라 선별적으로 처리하고자 하는 차후연구에 활용하기 위해서이다.

위의 조건에 의해서 분류된 뉴스 데이터는 수동으로 기사 단위로 분할한 다음 이를 문장 단위로 재분할하였다. 최종 데이터베이스의 세부사양은 [표 2]와 같다.

3. 뉴스 데이터베이스의 파일 인덱싱

차후 연구를 위해서 다양한 옵션을 적용하여 뉴스 데이터베이스를 처리하였다.

1단계는 방송국, 뉴스 방송 시간대이다. 2단계는 방송일자이다. 3단계는 화자정보를 바탕으로 앵커, 리포터, 시민, 복수화자, 기타로 인덱싱하였다. 이때 앵커, 리포터, 시민은 남녀를 구분하였으며, 각 화자에는 고유번호를 부여하였다. 뉴스 데이터베이스의 화자현황은 [표 3]과 같다. 시민의 수가 전체의 75.4%를 차지한다. 이러한 정보는 차후에 화자분류 및 화자추적 등을 통해서 뉴스 데이터를 좀 더 효율적으로 분석할 때 사용될 것이다. 4단계는 기사번호이다.

이런 규칙으로 만들어진 뉴스 데이터 인덱싱은 다음과 같다. "MM_1210_MA_001.wav"은 아침뉴스 MBC, 12월 10일, 남자 앵커, 기사번호 001을 뜻한다.

III. 뉴스 데이터 분석 : 전처리

1. 뉴스 데이터 분석 시스템

뉴스 데이터 분석 시스템[그림 1]은 4개의 핵심모듈로 구축될 예정이다. 각 핵심모듈은 독립적으로 실험을 진행하였으며, 현재 2번째 핵심모듈까지 연구가 진행되었다[5][6]. 본 논문은 3번째 핵심모듈인 연속음성 인식기에 대한 연구 결과이다.

뉴스 데이터 분석 시스템에 대해서 간단히 설명하면 다음과 같다. 무음구간 검출 & 분할까지는 뉴스 데이터베이스 구축에서 이루어진다. 1번째 핵심모듈인 음성 / 음악 분류기는 입력음향 중에서 음성 부분만을 선별한다. 2번째 핵심모듈인 화자변화 검출기는 음성부분을 화자의 변화를 기준으로 분할하여 단일 화자로 이루어진 음성 클러스터를 생산한다. 3번째 핵심모듈인 연속음성 인식기는 입력음성을 인식하고 그 결과를 후처리하여 핵심어 정보 파일을 생산한다. 4번째 핵심모듈인 뉴스 데이터 검색기는 핵심어 정보 파일을 분석하여 사용자의 질의에 맞는 뉴스를 응답한다.

뉴스 데이터 분석 시스템의 이해를 돕기 위해서 연속음성인식을 제외한 나머지 부분에 대해서 간략하게 설명한다.

2. 음성 / 음악 분류기

뉴스 데이터는 특성상 여러 가지 음향 정보가 혼재되어 있다. 이 중에서 사람의 음성만을 추출하는 것이 음성 / 음악 분류기의 역할이다. 이는 연속음성 인식에 적합한 음성만을 추출하여 전체 시스템의 안정성을 높이고 계산 비용을 절감하는데 목적이 있다.

음성 / 음악 분류를 위한 특징 파라미터는 멜 캡스트럼(Mel Cepstrum), 영교차(Zero Crossing) 관련 파라미터, 정규화 로그 에너지와 이들의 델타(Delta) 파라미터를 사용하였다.

음성 / 음악 분류를 위한 알고리듬은 GMM (Gaussian Mixture Model)을 사용하였다.

실험결과, 특징 파라미터 단독보다는 상호 조합하였을 때 좋은 결과를 나타내었다. 멜 캡스트럼, 에너지를 특징 파라미터를 상호 조합하여 실험한 결과 음성 98.9%, 음악 100%의 성능을 보였다[5].

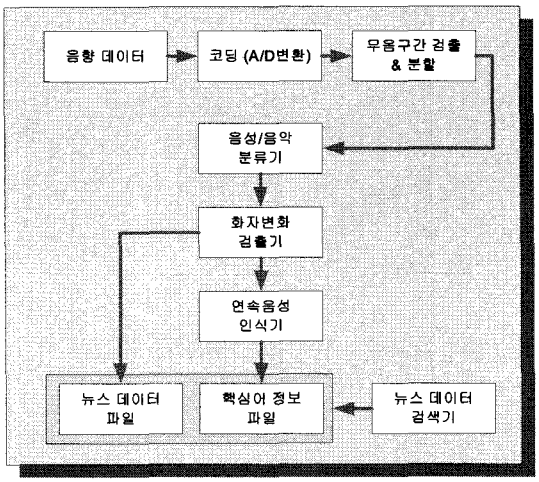

그림 1. 뉴스 데이터 검색 시스템

3. 화자 변화 검출기

뉴스 데이터는 "앵커-리포터-기타" 순으로 진행순서가 정형적이다. 그리고 앵커의 발성부에서 기사의 컨텐트에 대해서 언급된다. 또한 정확한 발성과 조용한 배경 음향 환경을 가지고 있어 음성인식을 위한 좋은 입력 데이터가 된다. 그러므로 화자정보는 뉴스 데이터 분석의 신뢰도 및 효율성을 높이는데 중요하다.

논문 [6]에서는 기사 단위로 만들어진 뉴스 데이터를 화자정보를 바탕으로 단일화자로 구성된 음성 클러스터로 분할하였다.

화자 변화 검출은 전송채널 및 입력신호의 변화에 민감한 GLR(Generalized likelihood ratio) 거리 기반 방법으로 화자 변화 지점을 추정한 다음, 긴 분석 윈도우가 확보되었을 때 강인한 BIC(Bayesian Information Criterion) 기반 방법으로 검증하였다.

실험결과 고소음 하에서는 MDR(Missed Detection Rate) 10.20%, FAR(False Alarm Rate) 80.93%, SR(Shift Rate) 0.83초의 성능을 보였고, 저소음 하에서는 MDR 6.25%, FAR 67.67%, SR 0.54초의 성능을 보였다[6].

4. 뉴스 데이터 검색기

뉴스 데이터 검색은 연속음성 인식기에서 만들어진 핵심어 정보 파일을 이용할 예정이다. 사용자가 검색어를 입력하면 핵심어 정보 파일을 검색하여 검색어가 포함된 뉴스 기사들을 응답한다. 검색결과가 복수 개일 경우에는 해당 검색어의 로그확률 순으로 내림차순으로 정렬한다.

Ⅳ. 뉴스 데이터 분석 : 연속음성인식

1. 연속음성인식을 위한 데이터베이스

데이터베이스는 음향모델을 훈련하기 위한 훈련 데이터베이스와 평가를 위한 평가 데이터베이스로 구성되었다.

훈련 데이터베이스는 51개의 뉴스 데이터 중에서 음성의 배경환경이 silence인 것만을 선별하여 구축하였다. 이는 음성인식 시스템의 기본인 음향 모델들을 훈련하는데 있어서 배경소음이 포함된 훈련 데이터를 사용하면 음향 모델이 열화 되어서 인식성능이 저하될 것을 우려해서다. 일반적으로는 음성인식이 적용되는 배경환경과 동일한 곳에서 훈련 데이터를 수집하고 이를 바탕으로 음향 모델을 모델링하는 것이 정석이겠지만, 뉴스 데이터의 특성상 그 배경환경이 매우 다양하여 모든 것을 다 고려한다는 것은 불가능하다. 수동으로 분류한 결과 총 21,059 문장 중에서 7,912 문장이 훈련 데이터로 선정되었다

훈련 데이터베이스는 배경소음 조건으로 인해서 발성자의 대다수가 앵커였다. 이는 화자독립 연속음성인식에 좋지 않은 결과이다. 차후 연구에서 이 부분에 대해서 개선이 필요할 것이다.

평가 데이터베이스는 전체 51개 뉴스 데이터 전체로 구성되었다.

2. 연속음성인식

뉴스 데이터는 자연스러운 발성으로 이루어져 있으므로 여러 음성인식 방법 중 연속음성인식이 가장 적절한 분석 수단이라고 할 수 있다.

연속음성인식을 위한 특징 파라미터로는 12차 멜 캡스트럼, 1차의 정규화 로그 에너지, 이들의 델타 파라미터를 사용하였다.

연속음성인식은 과거의 음성신호 처리만을 이용하는 방법에서 언어처리와 이해기법을 통합한 형태로 발전하였다. 특히, 한국어 연속음성인식의 경우에는 형태소 분석이 반드시 필요하다[4].

형태소는 최소 유의적 단위라고 정의된다. 형태소 분석이란 주어진 문장의 최소 의미 단위인 형태소를 추출하는 것이다. 전체 데이터베이스에서 음절 단위의 고유 개수를 살펴보면 77,749개나 되지만, 이를 형태소 분석하면 31,000개로 감소하고, 이것을 수동으로 검증하면 19,113개로 감소한다. 이러한 방법의 형태소 분석은 짧은 형태소를 다수 발생시켜 차후 인식의 성능에 좋지 않은 영향을 미친다. 하지만, 적절한 규모의 탐색공간을 확보하기 위해서 이를 감수하였다.

이렇게 분석된 형태소는 음향모델 생성을 위해서 음소 단위로 분할되었다. 이 과정에서 동일한 발성을 가지는 형태소들이 통합되어 최종적으로 18,993개의 형태소가 음성인식에 사용되었다.

연속음성인식은 HTK(Hidden Makov model Tool Kit)를 이용하여 구축된 음향 모델과 CMU(Carnegie Mellon University)의 SLM(Statistical Language Modeling) toolkit을 이용하여 구축된 언어모델을 사용하였다[7][8].

입력 음성은 특징 파라미터들의 수열로 변환되고, 이것은 토큰 패싱 알고리듬(Token Passing Algorithm)으로 구성된 인식 네트워크에서 음향 모델과 언어 모델을 참조하여 인식을 수행하게 된다.

참고논문 [9]에서는 뉴스 데이터에 대해서 음성 인식을 수행한 결과를 보여준다. 이때 사용된 음성인식 시스템은 CMU에서 개발한 대용량 화자독립 연속 음성 인식기인 Sphinix-Ⅱ이다.

실험결과, 음성인식 성능은 데이터의 품질과 성질에 따라서 급격한 성능변화를 보인다. 그 중 우리와 가장 유사한 형태인 30분 분량의 저녁 뉴스에 대한 인식결과를 살펴보면 WER(Word Error Rate)가 65%였다.

한국어 방송의 분석에 대한 국내연구를 살펴보면 다음과 같다. 음향모델의 훈련 데이터는 KBS 뉴스 16회분으로 구축되었으며, 언어모델의 훈련 데이터는 2년 반의

방송뉴스 원고(6.6M)와 2년의 뉴스기사(75M)로 구축되었다. 또한, 평가 데이터는 2회 분량의 방송을 선택하였다.

실험결과, 기본성능의 인식기가 앵커 발성부에서 22.8%의 WER, 리포터 발성부에서 34.3%의 WER를 보였다[10].

위의 결과는 언어모델 훈련 데이터베이스가 51회의 방송뉴스 원고(0.86M)로 이루어진 것과, 평가 데이터가 51회 분량의 방송이며, 앵커와 리포터의 발성부를 별도로 실험했음에도 불구하고 낮은 성능이다.

인식 네트워크를 이용한 인식과정은 다음과 같다. 음향모델의 입력으로 전파된 토큰이 음향모델의 출구 상태에 도달하면 언어모델을 참조하여 각 음향모델의 입력으로 재전파된다. 이러한 과정이 반복되면서 토큰에 음향모델의 출력 열이 저장된다.

이 과정에서 언어모델이 한국어 연속음성 인식의 성능을 저하하는데 영향을 미친다. 참고논문 [10]에서는 본 논문보다 94배나 많은 언어모델 훈련 데이터베이스를 사용하여 보완하려고 하였으나 낮은 성능을 보였다.

본 논문에서는 언어모델의 이러한 영향을 약화시키기 위하여 매 프레임별 음향모델의 출구 상태에 도달한 토큰들이 언어모델을 참조하여 각 음향모델의 입력으로 전파되기 전에 이를 출력하였다. 즉, 언어모델이 적용되기 전의 음향모델의 인식결과만을 출력한 것이다. 토큰의 출력은 로그확률을 이용하여 정렬하고 출력레벨 문턱치를 충족시키는 것만을 선별하였다.

[그림 2]는 본 논문에서 제안한 음향모델의 출력을 이용한 연속음성인식을 나타낸 것이다.

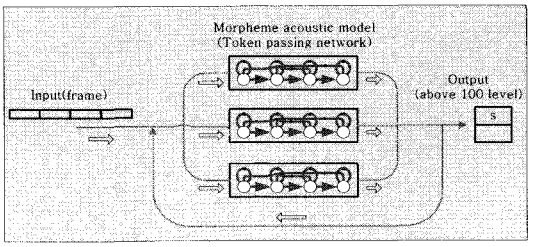

그림 2. 음향모델 출력을 이용한 연속음성 인식

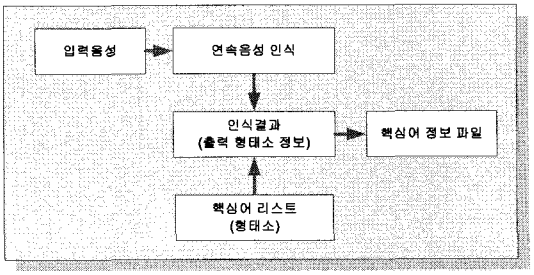

그림 3. 인식결과 후처리

표 4. 명사 계열 형태소 현황

구 분	명사 계열	비명사 계열
수 (비율)	13,144 (69.2%)	5,849 (30.8%)

V. 실험결과

1. 후처리

매 프레임별 연속음성인식의 음향모델 출력은 후처리 과정을 통해서 가장 확률이 높은 것들을 선별하여 핵심어 정보파일의 형태로 출력된다.

매 프레임별 각 음향 모델의 출력들은 핵심어(인식 대상 형태소) 리스트와 비교하여 핵심어 번호로 변환된다. 핵심어 번호로 변환하는 것은 핵심어의 길이에 따른 용량 변화를 방지하기 위해서이다. 핵심어 번호로 변화된 출력 형태소는 로그 확률을 기반으로 하여 내림차순으로 정리된다. 이 때, 복수 출현한 핵심어는 가장 높은 확률을 가지는 것만을 출력한다.

위의 작업은 입력 파일마다 실행되고 그 결과로 핵심어 번호와 로그 확률을 가진 핵심어 정보파일이 만들어진다.

2. 인식대상 형태소 선정

음향모델의 출력을 분석한 결과 문장의 핵심정보를 갖지 않는 것들이 상위레벨에 분포하는 것이 관찰되었다. 예를 들어 "북한의 금강산 관광이 재개되었다."에서 핵심정보는 "북한", "금강산 관광", "재개"이다. 즉, 명사 계열의 형태소가 문장의 핵심정보이다.

그래서 출력 형태소 중에서 명사 계열만을 인식대상으

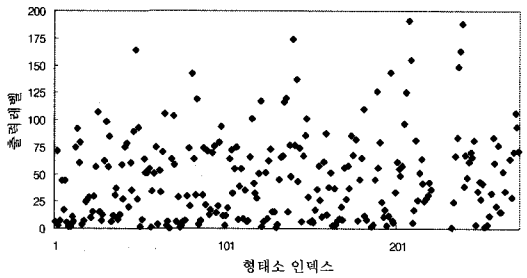

그림 4. 인식대상 형태소의 출력레벨

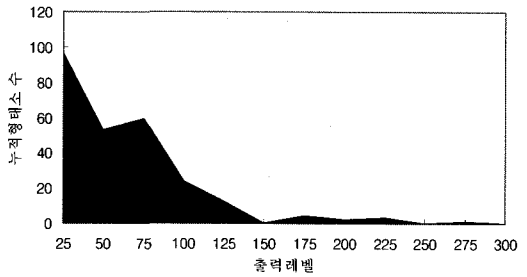

그림 5. 인식대상 형태소의 출력레벨별 분포

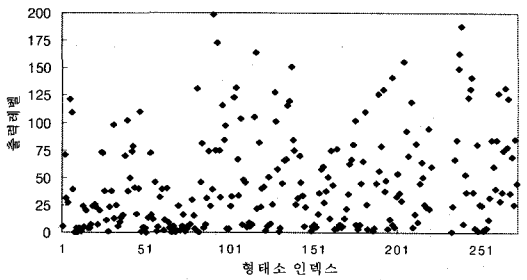

그림 6. 인식대상 형태소의 출력 레벨
(형태소의 길이정보로 정규화 후)

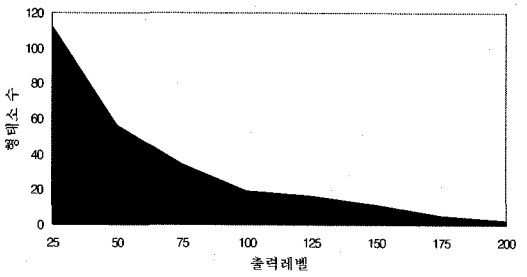

그림 7. 인식대상 형태소의 출력 레벨별 분포
(형태소의 길이정보로 정규화 후)

표 5. 출력레벨별 인식대상 형태소의 비율

구 분	50	75	100	125
점유율(%)	55.8	77.9	86.9	91.4

표 6. 출력레벨별 인식대상 형태소의 비율
(정규화 적용 후)

구 분	50	75	100	125
점유율(%)	61.76	74.26	81.25	87.13

로 받아들였다. 명사 계열에는 불완전명사, 동작성 명사, 보통명사, 시간성 명사, 상태성 명사, 인칭고유명사, 고유명사, 비단위성 의존명사, 성씨 고유명사, 비서술성 명사, 단위성 의존명사, 상태성 명사, 동작성 명사가 있다. 그리고 단음절 명사도 수동으로 정리하였다.

[표 4]는 전체 형태소 중에서 명사 계열의 현황을 나타낸 것이다.

3. 출력레벨 결정

매 프레임별 각 음향모델의 출력은 후처리 과정에서 로그확률을 기반으로 내림차순으로 정리된다. 이 때 프레임별 출력 레벨의 결정이 선행되어야 한다. 출력 레벨이 너무 낮아서 아주 미미한 확률을 가진 출력까지 모두 출력이 된다면 연산비용이 증가하고, 출력 결과 파일의 크기가 급격히 증가하여 전체 시스템에 대한 과도한 저장 공간 요구가 발생할 수 있다.

우선 30개의 파일에 대해서 매 프레임별 모든 출력을 받아들였다. 그리고 이들 중 우리가 원하는 출력들이 존재하는지를 살펴보았다.

[그림 4]는 출력 형태소 중 입력 문장과 동일한 출력, 즉, 바르게 인식된 형태소들의 출력레벨 분포를 나타낸 것이다. 이를 통해서 목적하는 전체 인식대상 형태소의 80% 이상을 처리할 수 있는 출력레벨을 결정하고자 하였다.

[그림 5]는 각 출력레벨별 인식대상 형태소의 분포를 나타낸 것이다. 낮은 출력레벨(즉, 확률이 높은 레벨)에 대부분의 인식대상 형태소가 몰려 있고, 높은 출력 레벨(즉, 확률이 낮은 레벨)에는 소수의 인식대상 형태소가 존재한다. [표 5]는 출력레벨을 변화시켰을 때 출력 영역

에 포함되는 인식 대상 형태소의 비율을 나타낸 것이다. 100 레벨 정도에서 전체 인식 형태소의 86.9%가 포함되는 것을 알 수 있다.

4. 형태소 길이정보를 이용한 정규화

위의 결과는 음향모델의 길이정보를 고려하지 않은 것이다. 일반적으로 음향모델의 길이가 길어지면 인식 네트워크의 길이가 길어지면서 출구 상태에 도달한 토큰의 확률이 작아지게 된다. 즉, 길이가 긴 형태소는 출력 로그확률이 작고, 짧은 길이의 형태소는 상대적으로 크다는 것을 고려하지 않은 것이다.

이를 해결하기 위해서 본 논문에서는 길이정보를 해당 음향모델에 입력되는 프레임의 수로 정의하고, 이를 이용하여 각 음향모델의 출력들을 정규화 하였다.

[그림 6]은 각 형태소의 길이 정보로 출력확률에 정규화를 적용한 실험에서 인식대상 형태소의 출력레벨 분포를 나타낸 것이다.

[그림 7]은 각 형태소의 길이 정보로 출력확률에 정규화를 적용한 실험에서의 각 레벨별 인식대상 형태소의 수를 나타낸 것이다.

형태소의 길이정보를 이용하여 정규화한 결과 [표 6]에서와 같이 출력 레벨이 100일 때 정규화를 적용하기 전에는 86.9%였던 인식률이 81.25% 로 저하되었다. 이러한 성능저하에도 불구하고 길이정보를 이용한 정규화는 필요하다.

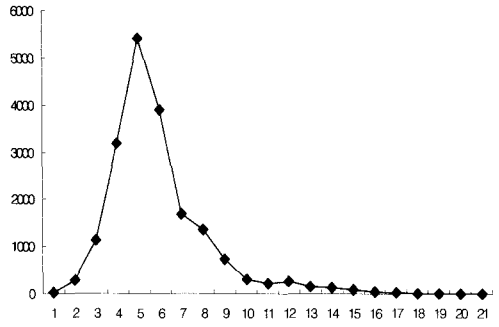

그림 8. 전체 형태소의 구성음소수 분포

표 7. 원본 DB와 핵심어 정보 파일 DB

구 분	파일 수	크기(MB)
원본 DB	2,093	5,040
핵심어 정보 파일 DB	2,093	314

이를 위해서 길이정보와 형태소를 구성하는 음소 수가 비례한다는 것을 감안하여 전체 형태소의 구성 음소수를 조사하였다. 그림 8은 전체 18,993개의 형태소의 구성 음소수의 분포를 나타낸 것이다. 상대적으로 짧은 길이인 5개의 음소 이하로 이루어진 형태소가 전체의 52.9%를 차지하고, 10개의 음소 이하로 이루어진 형태소는 전체의 95.2%를 차지한다. 즉, 절반 이상의 인식 대상 형태소가 짧은 길이를 가진다는 것이다. 그러므로 전체 인식 대상 형태소 중 절반 정도를 차지하는 긴 길이의 형태소에 대해서 정규화가 필요하다.

이에 본 논문에서는 성능저하를 감수하고 길이정보 기반 정규화를 실험에 적용하였다.

5. 핵심어 정보파일 DB 구축

음향모델의 출력레벨 문턱치를 100으로 설정하고 길이정보 기반 정규화를 적용하여 실제 뉴스 데이터베이스에 대하여 실험하였다.

본 논문에서 사용한 뉴스 데이터베이스는 51회분의 공중파 방송으로 이루어져 있으며, 파일형식은 8 kHZ 16 bit PCM(Pulse Code Modulation)이다.

실험결과 2,093개의 기사에서 전체 인식대상 형태소의 인식률은 75.13%였다. 참고논문 [10]과 비교하면, 앵커와 리포터의 발화부를 분리하지 않고, 입력음향의 품질이 낮은 시민의 발화부까지 포함하였으며, 51회의 방송분에 대해서 실험한 결과이다. 낮은 인식률을 보인 결과들을 분석한 결과, 짧은 길이의 형태소와 낮은 입력음향 품질(잡음, 실외잡음상태에서의 인터뷰 등)에 의한 인식성능의 저하가 주원인이었다. 이는 차후 연구에서 보완해야 할 사항이다.

가공하기 전의 뉴스 데이터베이스는 [표 7]에서 보는 바와 같이 2,093개의 기사로 이루어졌으며 용량은 5,040 MB였다. 이것을 길이정보 기반 정규화를 적용하여 실험한 결과 314 MB 용량의 핵심어 정보 파일이 생산되었다. 절대적인 정보량에서 기존의 원본 데이터 대비 93.8% 감소하였다.

Ⅵ. 결 론

본 논문에서는 뉴스 데이터베이스 구축과 뉴스 데이터의 분석에 대해서 서술하였다. 실험에 사용한 뉴스 데이터베이스는 공중파 방송 51회분, 2,093개의 기사로 구성되었다. 분석도구는 한국어의 특징을 반영한 형태소 기반 연속음성 인식을 사용하였다. 연속음성 인식의 낮은 인식 성능을 고려하여 매 프레임별 각 음향모델의 출력을 인식결과로 받아들였다. 문장의 핵심정보가 집중되어 있는 명사 계열의 형태소만을 인식대상 형태소로 인정하였다.

음향모델의 출력 중에서 출력레벨이 100 이상인 것만을 유효한 출력으로 인정한 결과 전체 인식대상 형태소의 86.9%가 인식되었다.

짧은 길이의 형태소(5개 음소 이하)가 전체 인식대상 형태소 중 52.9%를 차지하는 사실을 반영하여 길이정보 기반 정규화를 적용하였다. 실험결과 인식대상 형태소의 포함률이 81.25%로 약간 저하되었다. 본 논문에서는 긴 길이의 형태소의 로그확률 보상을 위해서 길이정보 기반 정규화를 채택하였다.

위의 결과를 바탕으로 뉴스 데이터베이스를 분석할 결과 인식대상 형태소의 인식률은 75.13%였다. 그리고 5,040MB의 원본 데이터베이스에서 314MB의 핵심어 정보 파일을 얻을 수 있었다. 이것은 절대적인 정보량이 93.8% 감소한 것이다. 이를 통해서 뉴스 검색 시 검색공간의 축소와 검색비용의 절감을 기대할 수 있다.

참 고 문 헌

[1] J. C. R. Licklider, *Libraries of the future*, Cambridge, the MIT press, 1965.

[2] F. de Jong and T. Westerveld, "MUMIS : multimedia indexing and searching," Proc. of the content-based multimedia indexing workshop, pp.423-425, 2001.

[3] D. S. Pallet, "Overview of the 1997 DRAPA speech recognition workshop," Proc. 1997 DARPA Speech Recognition Workshop, Feb. 1997.

[4] O. W. Kwon, "Performance of LVCSR with morpheme-based and syllable-based recognition unit," Proc. ICASSP 2000, pp.1567-1570, 2000(6).

[5] 이경록, 서봉수, 김진영, "오디오 인덱싱을 위한 음성/음악 분류 특징 비교", 한국음향학회, 제20권, 제2호, pp.10-15, 2001.

[6] 이경록, 김진영, "통계적 기법을 이용한 화자변화 검출 실험", 음성과학, 제8권, 제4호, pp.59-72, 2001.

[7] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povery, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Dept, 2002.

[8] http://www.speech.cs.cmu.edu/SLM/toolkit.html

[9] H. D. Wactlar, A. G. Hauptmann, and M. J. Witbrock, "Informedia : News-on-demand experiments in speech recognition," DARPA speech recognition workshop, pp.56-60, 1996.

[10] O. W. Kwon and A. Waibel, "Korean broadcast news transcription using morpheme-based recognition units," The Journal of the Acoustical Society of Korea, Vol.21, No.1E, pp.3-11, 2002(3).

저 자 소 개

이 경 록(Kyong-Rok Lee) 정회원

- 1997년 2월 : 호남대학교 전자공학과(공학사)
- 2001년 8월 : 전남대학교 정보통신협동과정(공학석사)
- 2006년 2월 : 전남대학교(공학박사)
- 2003년 3월~현재 : 남부대학교 디지털정보학과 교수

<관심분야> : 음성인식, 멀티미디어 검색