운율 분석용 DB 작성을 위한 자동 레이블러(Automatic labeler)의 성능 평가 및 유용성

김상훈, 이항섭, 김회린

305-600 대전광역시 유성우체국 사서함 106 호
한국전자통신연구소 휴먼인터페이스부 음성언어연구실
Tel: 042-828-6234, Fax: 042-828-6231

요약

이 논문에서는 대량의 음성합성용 운율 DB 를 용이하게 구축하기 위해 음성번역시스템을 이용한 자동 레이블러의 성능을 다양한 음성데이타를 대상으로 평가하였다. 실험 결과 FM radio news 문장, 대화체 문장 및 낭독체 문장 등에는 레이블링 대상 음소의 약 80% 이상이 오류가 30msec 이내인 범위로 레이블링 되며, 고립단어에 대해서는 약 60%의 성능을 보여주고 있다. 현재 당 연구실에서는 자동 레이블러를 이용하여 합성용 운율 DB 및 합성단위를 작성하고 있으며, 자동 레이블러를 이용함으로서 일관성 있는 레이블링 결과를 얻을 수 있을 뿐 아니라 작성하는데 소요되는 시간도 줄일 수 있었다.

1. 서론

합성용 음성 DB 의 구축은 음소단위의 분절(segmentation)과 레이블링(labeling)이 필수적으로 요구되며, 이러한 작업은 단순하면서도 시간이 많이 소요된다. 특히 최근에는 음성합성을 위한 운율분석이 대량의 음성데이타를 기반으로 이루어질 뿐만 아니라 다양한 화자가 발성한 합성단위 및 새로운 합성단위의 구축이 필요함에 따라 자동 레이블링의 필요성이 점점 중요해지고 있다. 따라서 이 논문에서는 당 연구실에서 개발한 음성번역시스템을 이용하여 고립단어, 낭독체 문장, 대화체 문장, FM radio news 등 다양한 환경의 음성데이타를 대상으로 음소단위의 자동 분절과 레이블링에 관한 자동 레이블러의 성능을 기술하고 그 유용성에 대해 논한다.

2. 대화체 음성번역시스템 개요

당 연구실에서 1995 년도에 개발한 음성언어 번역시스템은 대화체 음성인식 및 한/일, 한/영 번역을 목적으로 하고 있다. 현재 여행영역에서 약 5,000 단어를 인식할 수 있으며, HMM 인식 알고리즘을 사용하고 한국어 음운환경을 고려한 allophone clustering tree 를 이용하여 단어 인식률이 약 70%(perplexity=110)에 이르고 있다[1][2].

3. 음성 데이타베이스

성능평가를 위해 사용된 음성 DB 는 고립단어(3848 POW), 낭독체 문장, 대화체 문장(통역자), 대화체 문장(대화자), FM Radio News 로 구성되어 있고, 모두 16KHz, 16bit sampling 되어 있다. 그리고 각 음성 DB 는 음성학 전문가 및 음성합성 분야 전문가에 의해 수동으로 레이블링이 되어 있다. 표 1 은 성능평가에 사용된 음성 DB 들의 구축 목적 및 사양을 설명하고 있다[3].

표 1. 음성 DB 의 종류

음성 DB	구축 목적 및 사양
고립단어 (3848 POW)	남녀 각 8 명이 3848 POW 을 나누어 발성한 POW 2 set. 고립단어 음성인식 훈련용 데이타.
낭독체 문장	낭독체 음성합성의 운율모델링용 음성데이타. 1 명의 여성 아나운서가 발성한 156 개 문장으로 구성
대화체 문장 (통역자)	대화체 음성합성의 운율모델링용 데이타. 1 명의 남성통역자가 발화한 64 개 utterance 로 구성
대화체 문장 (대화자)	대화체 음성합성의 운율모델링용 데이타. 남성 2 명이 발화한 111 개 utterance 로 구성
FM Radio News	낭독체 음성합성의 운율모델링용 음성데이타. FM radio news 의 남녀 아나운서 각 29 개 문장으로 구성

고립단어에 대한 성능평가 목적은 새로운 합성

단위의 구축시 고립단어의 분절이 필요함에 따른 것이며, FM radio news는 전문 아나운서의 발성음을 용이하게 녹취할 수 있어 자동 레이블링이 FM radio 대해서도 유효하다면 대량의 음성 데이타를 손쉽게 구축할 수 있다.

4. 분절 및 레이블링 과정

자동 레이블링의 첫 단계는 발성한 음성데이타를 텍스트로 적는 전사(transcription) 과정이다. 이때 전사는 grapheme으로 표기하고 전사된 한글표기는 발음사전과 utterance file 작성을 위해 영문표기로 변환한다. 두번째 단계에서는 음성데이타에 포함된 단어(띄어쓰기 단위 또는 어절)를 grapheme-to-phoneme 변환기를 사용하여 소리나는대로 바꾸어 발음사전을 작성한다. 발음사전에는 단어 이외 묵음, 입술소리, 기침소리, 기타 잡음 등의 기호가 포함되어 있다. 다음 단계에서는 레이블링할 음성데이타를 문장단위로 나누고 이 문장을 영문표기로 바꾸어 utterance file을 작성한다. Utterance file은 전사된 텍스트 및 그 문장에 해당하는 음성의 feature file의 위치 정보를 담고 있다. 마지막으로 레이블링할 음성데이타로부터 feature file을 생성하고 음성번역기를 수행한다. 레이블링에 사용된 feature는 청각적 영향을 고려한 FFT-based LPC를 사용한다. 자동 레이블링 전체 과정은 그림 1과 같다.

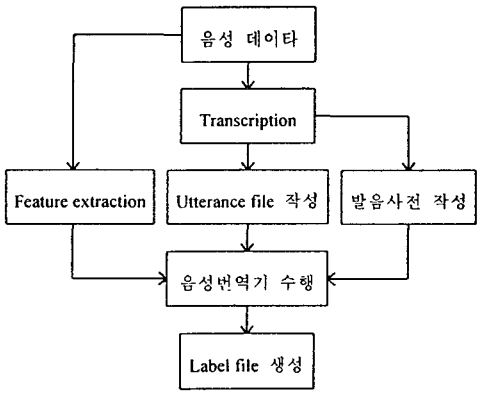


그림 1. 자동 레이블링 과정

그림 2는 자동 레이블링한 결과를 보여주고 있다. 레이블링에 사용된 음소기호는 모음 19개, 초성자음 18개, 종성자음 7개로 정의하였다.

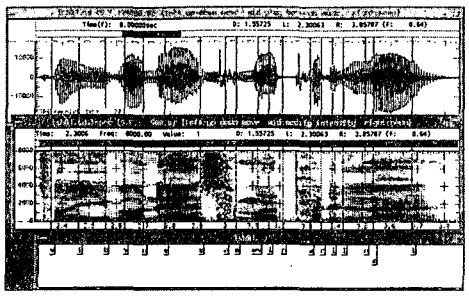


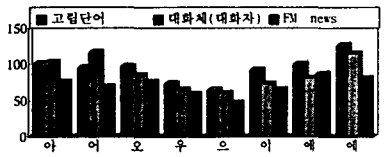
그림 2. 대화체 음성의 자동 레이블링 결과: "저 이번에 삼월 팔일날"

5. 결과 및 분석

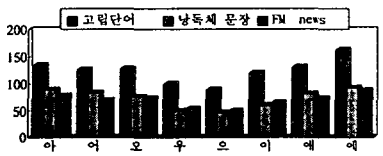
성능 비교를 위하여 수동으로 레이블링한 결과와 비교하였으며 각 음성 DB에 대한 성능은 아래의 표 2와 같다. 비교는 수동에 의한 레이블링 위치와 자동에 의한 레이블링 위치를 비교하여 그 차이의 절대치를 E (*Error* = | *hand label - auto label* |)로 하였다. 그리고 자동 레이블링과 수동 레이블링된 결과사이에 음소의 삽입, 생략 및 대치가 발생하는데 이는 전체 음소 갯수에 비해서 적은 양이므로 여기서는 무시한다. 주로 생략 오류는 종성 /ㄱ,ㄷ,ㅂ/ 및 유성음 사이에서의 /ㅎ/ 음소이며, 삽입의 오류는 /ㄴ/, /ㄹ/ 음소에서, 대치의 오류는 [관광->광광], [입니다->임미다]와 같이 /ㄴ/음소가 /ㅇ/ 또는 /ㅁ/으로 동화되어 발생한다.

자동 레이블링 오류가 30 msec 이내인 결과를 고려할때, 고립단어인 경우 남녀 각 62% 63%, 낭독체 문장인 경우 82%, 대화체 문장의 통역자인 경우 87%, 대화체의 대화자인 경우 78%, FM radio news의 남녀 모두 87%의 성능을 보여주고 있다. 고립단어의 자동 레이블링 결과가 기타 문장 음성 DB보다 성능이 저하되는 주요 원인으로는 대화체 scheduling domain에서 발생하는 triphone(5,800여개)이 고립단어 즉 3,848개의 POW에서 발생하는 triphone (9,000여개)을 포함하지 못한 데 있고 다음으로는 레이블링에 사용된 codebook이 대화체 문장으로 훈련되어 고립단어에서의 음성신호 특성을 포함하지 못한데 있다. 즉 각 음성 DB 별 지속시간 분포로 음성신호의 변화를 간접적으로 분석하면, 그림 3에

서와 같이 낭독체문장, 대화체문장, FM radio news 문장의 단모음에 대한 지속시간 분포가 대화체에서의 간투사에 해당하는 단모음 /아.어/를 제외하면 서로 유사한 값을 가지고 있으나, 고립단어와는 상당히 다르다. 따라서 지속시간의 변화로부터 알 수 있는 사실은 문장음성 DB의 경우 고립단어로 발성된 음성보다 동시조음의 영향이 커 서로 다른 음성신호특성을 가짐을 알 수 있으며, 고립단어의 경우 음소간 천이구간이 길어 40msec 이상 오류도 높게 나타난 것으로 분석된다.



가. 남성음



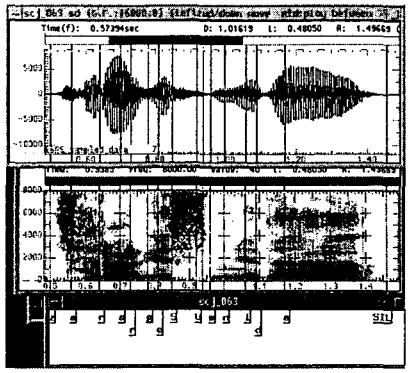
나. 여성음

그림 3. 음성 DB에 따른 단모음 지속시간 분포

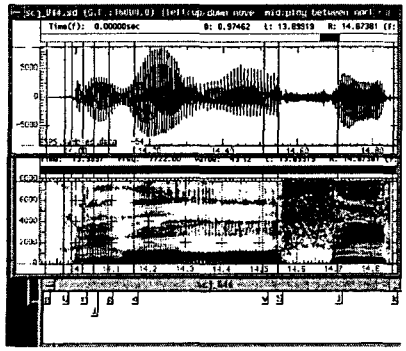
또한 대화체 대화자 음성의 경우 전사 오류로 인해 40msec 이상 레이블링 오류가 대화체의 통역자 문장, 낭독체 문장, FM radio news 문장 음성보다 약 2배 정도 높아 전체적으로 성능이 다소 떨어진 결과를 보여주고 있다. 각 음소별 category에 따른 성능에서는 모음보다 자음의 레이블링 정확률이 높게 나타났으며, 복모음보다는 모음의 레이블링 정확률이 높게 나타났다 (표 3, 표 4). 특히 비음부에서는 상당히 정확한 레이블링이 되었으나, 그림 4에서와 같이 '모음+모음, 모음+유음, 파열/파찰음+모음'의 경우 음소 경계에서 오류가 크게 나타났다. 성별에 따른 성능은 여성음이 남성음보다 1~2%정도 저하되는 것을 알 수 있다.

6. 결론

대화체로 훈련된 파라미터를 사용함으로서 고립단어의 경우 상당한 오류를 보이고 있으나, 문장 DB에서는 자동 레이블러가 유용함을 알 수 있다. 위의 결과로부터 자동 레이블러의 오류를 최소화하기 위해 고립단어, 문장 DB 남녀에 대



가. '모음+유음, 파열음+모음'인 경우



나. '모음+모음'인 경우

그림 4. 자동 레이블링 오류 예

한 파라미터 작성이 필요함을 알 수 있다. 실제 당 연구실에서는 자동 레이블러를 이용하여 음성합성용 운율 DB를 작성하고 있으며, 과거 수동 레이블링에 비해 시간과 노력을 상당히 줄일 수 있었다. 또한 자동 레이블러를 사용함으로써 수동 레이블링에 의한 레이블 기호의 생략 및 대치를 없앨 수 있고 레이블 위치에 대한 수정이 간단해지며, 레이블링의 일관성을 얻을 수 있다. 특히 FM radio는 전문 아나운서의 음성 녹취가 용이하고 자동 레이블러를 이용하여 대량의 운율 DB를 빠른 시간내에 구축할 수 있는 장점이 있다. FM radio는 이미 외국에서는 널리 이용하는 방법이나 국내에서는 아직 미흡하므로

이를 이용한 운율 DB 구축이 이루어져야 할 것이다[4][5].

참고문헌

[1] 이영직, 양재우,"다중매체 통신을 이용한 대화체 음성언어번역 시스템", 제 13 회 음성통신 및 신호처리 워크샵, pp 101-106, 1996.

[2] 김회린, 이항섭,"POW 3848 단어 인식기 구현 및 어휘독립 실험", 제 13 회 음성통신 및 신호처리 워크샵, pp 127-130, 1996.

[3] 이영직 외," ETRI 의 음성 데이타베이스 구축 현황", 제 12 회 음성통신 및 신호처리 워크샵, pp 265-267, 1995.

[4] F.Emerard, L.Mortamet,"Prosodic processing in a text-to-speech synthesis system using a database and learning procedures," in Talking Machines:Theories, Models, and Designs, North-Holland, pp.225-254, 1992.

[5] Jung-Chul Lee, Sanghun Kim and Minsoo Hahn, "Intonation processing for Korean TTS Conversion Using Stylization Method," in Proc. ICSPAT'95,

표 2. 음성 DB 에 따른 자동 레이블러의 성능

음성 DB	고립단어		낭독체 문장	대화체 문장 (통역자)	대화체 문장 (대화자)	FM Radio News	
성별	남	여	여	남	남	남	여
음소 갯수(개)	24,290	23,790	14,450	5,127	3,772	4,053	3,914
E<=10msec	29%	28%	41%	58%	46%	58%	57%
10msec <E <= 20msec	20%	21%	26%	19%	19%	19%	21%
20msec <E <= 30msec	13%	14%	15%	10%	13%	10%	9%
30msec < E<= 40msec	13%	13%	7%	5%	7%	5%	3%
40msec <E	25%	25%	11%	8%	14%	7%	9%

표 3. 대화체 문장의 음소 category 에 따른 레이블러의 성능

음성 DB	대화체 문장(대화자)			
성별	남			
음소 category	모음	복모음	초성자음	종성자음
음소 갯수(개)	1,552	280	1,392	548
E<=10msec	41%	39%	46%	64%
10msec < E <= 20msec	20%	16%	21%	16%
20msec <E <= 30msec	16%	11%	12%	8%
30msec < E <= 40msec	8%	11%	6%	4%
40msec < E	15%	23%	14%	8%

표 4. FM radio news 문장의 음소 category 에 따른 레이블러의 성능

음성 DB	FM Radio News							
성별	남				여			
음소 category	모음	복모음	초성자음	종성자음	모음	복모음	초성자음	종성자음
음소 갯수(개)	1,581	239	1,580	653	1,537	220	1,521	636
E<=10msec	49%	52%	64%	70%	53%	57%	59%	60%
10msec < E<= 20msec	23%	21%	17%	16%	24%	16%	20%	18%
20msec < E <= 30msec	14%	13%	7%	4%	11%	11%	8%	9%
30msec < E <= 40msec	7%	5%	4%	4%	4%	5%	3%	4%
40msec < E	7%	9%	8%	7%	8%	11%	10%	9%