음향 및 음소 정보를 이용한 연속체의 자동 음소 분할에 대한 연구

A Study on Automatic Phoneme Segmentation of Continuous Speech Using Acoustic and Phonetic Information

박 은 영*, 김 상 훈**, 정 재 호*
(Eun Young Park*, Sang Hun Kim**, Jae Ho Chung*)

※ 본 논문은 98년도 인하대학교 연구비 지원에 의하여 수행되었습니다.

요 약

본 논문은 자동 음소 분할기의 음소 경계 오류를 보상하기 위한 후처리(Postprocessing)에 관한 연구이다. 자동 분절 경계의 오류 범위를 줄일 수 있는 후처리기를 제안하고, 자동 분절 결과를 직접 합성 단위로 사용할 수 있는 대량의 합성용 운율 데이터 베이스 구축에 유용함을 기술한다.

제안된 후처리기는 수작업으로 보정된 데이터의 특징벡터를 다층 신경회로망(MLP: Multi-layer perceptron)을 통해 학습을 한 후, 자동 분절 결과와 MLP 기반 후처리를 이용하여 새로운 음소 경계를 추출한다. 우선, 특징벡터 set은 음성학적 지식이 최대한 반영되도록 선정되었다. 그리고, 경계를 추출하기 위해서 비선형 패턴분리에 탁월한 성능을 보이는 MLP를 이용한다. MLP는 매우 다양하게 나타나는 음소 경계간 음성학적 특징을 단시간 내에 적용할 수 있기 때문이다. 마지막으로, 음운환경별로 특징 벡터가 적용되는 제안된 후처리 알고리즘을 이용하여 자동 분절의 경계 오류에 대한 보상이 이루어 진다.

문장 단위로 발화된 합성용 데이터베이스에서 후처리기로 보정된 분절 결과는 음성 언어 번역 시스템의 분할율보다 약 19.9%의 향상된 성능을 보였으며, 절대오류 (|Hand label position-Auto label position|)는 약 28.6% 감소되었다.

ABSTRACT

The work presented in this paper is about a postprocessor, which improves the performance of automatic speech segmentation system by correcting the phoneme boundary errors. We propose a postprocessor that reduces the range of errors in the auto labeled results that are ready to be used directly as synthesis unit.

Starting from a baseline automatic segmentation system, our proposed postprocessor trains the features of hand labeled results using multi-layer perceptron(MLP) algorithm. Then, the auto labeled result combined with MLP postprocessor determines the new phoneme boundary. The details are as following.

First, we select the feature sets of speech, based on the acoustic phonetic knowledge. And then we have adopted the MLP as pattern classifier because of its excellent nonlinear discrimination capability. Moreover, it is easy for MLP to reflect fully the various types of acoustic features appearing at the phoneme boundaries within a short time. At the last procedure, an appropriate feature set analyzed about each phonetic event is applied to our proposed postprocessor to compensate the phoneme boundary error.

For phonetically rich sentences data, we have achieved 19.9 % improvement for the frame accuracy, comparing with the performance of plain automatic labeling system. Also, we could reduce the absolute error rate about 28.6 %.

I. 서 론

음성은 인간의 가장 중요한 정보 교류 매체일 뿐 아니라, 인간과 기계 사이의 정보교환을 위한 MMI(Man-Machine Interface)에서 더욱 중요한 의미를 가진다. 최근 음성 신호

* 인하대학교 전자공학과, 디지털 신호처리 연구실
** 한국전자통신연구원, 통신단말 연구부

접수일자 : 1999년 2월 11일

처리 기술이 발달함에 따라, 제한적이나마 음성 인식, 음성 합성 기술을 바탕으로 여러 가지 상용 시스템이 개발되고 있으며, 특히, 음성 합성 기술은 음성 정보 서비스의 활성화와 수요의 증가로 더욱 중요한 연구 분야가 되고 있다. 고품질의 음성합성 system을 위해서는 합성음의 명료성과 자연성 모두가 개선되어야 하며, 이를 위해서 자연스러운 운율 부가를 위한 심층 연구가 진행되고 있다. 기존의 합성 시스템의 성능은 언어처리, 운율처리, 합성 방식, 합성단위 선정 등 상당히 방대하고 복잡한 요소에 의존한다. 자연스런 합성음을 위해서 주어진 입력(text)에 대해 적절한 prosodic contour(fundamental frequency, duration and amplitude)를 계산하고, 효과적으로 운율을 부여하는 등 신호처리 알고리듬 개발에 주력하였다. 그러나 운율 등의 무리한 규칙화와, 과도한 신호처리(signal processing)로 인한 왜곡(distortion)으로 기계음적인 부자연성을 극복하기 어려운 실정이다. 이러한 문제를 해결하기 위해, 신호처리를 최대한 감소시키는 대신 대량의 합성 DB로부터 가장 적합한 단위를 선택하여 합성하는 방식이 최근 활발히 연구되고 있다. 특히, 낭독체로 발성된 문장에서 추출하여 운율 현상이 충분히 포함된 합성단위를 사용하거나, 환경적 효과(contextual effect) 등 상관관계를 고려하면서 양질의 합성음을 얻을 수 있었다. 또한 원래 화자의 voice quality, speaking-style 특성을 그대로 간직하는 이점으로 음성 DB 구축에서 합성음 생성까지 자동 수행할 수 있는 연구가 지속적으로 진행되고 있다[1][2].

본 논문은 이러한 연구의 일환으로서, 풍부한 합성단위와 운율적 요소가 포함된 문장 단위 합성 DB구축과 나아가 합성단위 구축의 자동화 방법을 제시한다. 즉, 자동분절 오류를 간단히 수정할 수 있는 후처리 시스템을 제안하여 궁극적으로 합성음질의 향상을 목적으로 한다.

이에 앞서, 문장에 비해 스펙트럼 판독이 비교적 명확하고, 복잡한 조음현상의 발생빈도가 적은 어절단위 데이터를 음성시료로 하여 제안된 후처리 system을 적용하였다. 이때 비교적 간단한 단일 특징벡터를 이용한 결과 우수한 분절 성능을 얻을 수 있었다[3]. 본 논문은 위의 실험 결과 및 분석을 바탕으로, 다소 음소경계 검출이 어려운 문장단위에 적용하고자 한다. 특히, 제안된 방식을 단일 특징벡터가 아닌 복합적인 특징 벡터들을 적용하였을 때, 프레임 분할성능에 미치는 영향과 합성단위 자동생성에서의 유용성을 기술할 것이다.

본 논문의 구성은 다음과 같다. 2절에서는 음성 합성기를 구현하기 위한 합성 데이터 베이스 구축 과정 및 그의 개선점에 따른 현재 연구 동향을 기술한다. 3절은 제안된 자동 분할기의 오류를 수정하기 위한 후처리 방식이 자세히 기술된다. 마지막으로 4절 및 5절은 실험 결과와 결론에 따른 향후 연구 방향을 제시하고 마무리 짓는다.

II. 합성용 데이터베이스 구축

음성 합성 분야에서는 발음이 명료한 화자가 자연스럽게 발성한 대량의 데이터가 필요하며, 이를 발아 적은 녹취 데이터, 단어 및 음절, 음소 별로 나눈 레이블링 데이터가 절대적으로 필요하다. 또한, 운율 법칙 추출 등의 연구를 위해서는 피치 표시, 음소 단위 레이블링 작업이 요구된다. 기존 합성 데이터베이스 구축은 다음과 같다. 우선, 자동 레이블링 시스템을 이용하여 음소 경계를 분할하고, 음성 전문가에 의해 수작업으로 경계 보정이 이루어진다. 즉, 낭독체의 발음 특성을 고려하면서 정확한 음소 위치를 찾아내야 하므로 음성/음향학적 전문적인 지식이 바탕이 되어야 한다. 이러한 레이블링 작업은 기준에 따른 스펙트로그램 판독과, 반복되는 듣기평가를 통하여 이루어진다.

앞서 기술된 기존 방식과 더불어 최근 합성 데이터 베이스 구축 및 단위 선택에서 고려되고 있는 국내외 동향을 살펴 보면 크게 몇 가지 부류로 나눌 수 있다[1][2][4][5].

첫째, 음운 환경이 최대한 반영되거나 합성 단위간 연결부를 줄일 수 있도록 합성 단위가 점차 커지고 있다. 이는 합성 단위의 다양함과 선택의 중요성을 나타내는 것이다. 국내외의 경우 1997년 주변 음운환경을 고려한 음절 단위인 CDS(Context Dependent Syllable)뿐 아니라 여러 형태의 합성 단위가 경우에 따라 선택되는 방법, COC(Context Oriented Clustering) 방법 등이 연구 개발 중이다.

둘째, 합성 단위 연결시 복수 합성 단위 중 가장 적절한 합성 단위를 선택하여 접합하는 방식이다. 또한, 국내외 연구 동향에서 알려진 것과 같이, 음운 환경 뿐 아니라 피치, 지속 시간, 에너지 등 운율적 요소까지 고려된 합성 단위를 사용하려는 시도가 이루어 지고 있다. 셋째, 위에서 언급한 첫째, 둘째 방식이 효율적으로 이뤄지기 위해서는 대량의 합성 데이터 베이스 구축이 선행되어야 한다.

이를 위해 합성 DB 구축의 자동화가 절실히 요구되면서, 음성 인식기를 이용한 자동 음소 분할 및 레이블링 기술은 더욱 중요한 역할을 한다. 그러나 인식 및 레이블링 성능은 아직 만족할 만한 수준이 아니며, 양질의 합성음 생성을 위해 수작업에 의한 음소 경계 보정이 불가피하다.

그러나, 과거에 비해 DB 용량이 커짐에 따라 수동 분절 작업이 가지는 문제점은 더욱 크게 대두되고 있다[2][6]. 우선, 몇 사람의 전문가가 해오던 수작업에 의한 합성단위 분절은 상당기간(최고 4~5개월)이 소요되면서, 새로운 목소리의 합성음을 만들거나, 합성음을 사용하고자 하는 목적이 변경될 때 비생산적인 구축 방식이 된다. 또한, 음소 경계 기준을 미리 정해 놓더라도 여러 사람이 작업할 경우 일관성이 보장되지 못한다. 비록 정확도에서 자동 분절 결과가 수동 분절 결과에 비해 떨어지더라도 일관성이 유지된다면, 합성 단위간 연결점에서의 왜곡은 최소화 될 수 있을 것이다. 이에 본 논문은 수동 분절결과 분석을 이용한 후처리 방식으로, 자동 레이블링 시스템의 성능 향상에 대한 연구의 일환이다.

본 논문에서 사용된 자동분할기는 음성언어 번역 시스템 (The Spoken Language Translation System, ETRI)이다. 이의 분할 성능은 음소별, 음운 환경에 따라 성능 차이를 보이며, 음소 경계의 오류가 일정한 방향성을 가지면서

분절되는 것을 알 수 있다[3][7]. 따라서 이렇게 한쪽으로 치우친 자동 분절 결과를 그대로 이용한 합성음은 경계에서의 불연속(segment discontinuity)으로 인한 스펙트럼 왜곡이 발생하여 합성음의 음질이 떨어진다. 또한, 합성 단위로 사용되기에는 많은 오류를 포함한다. 그림 1은 자동 레이블링 시스템의 분할 오류의 구체적인 예를 나타내었다.

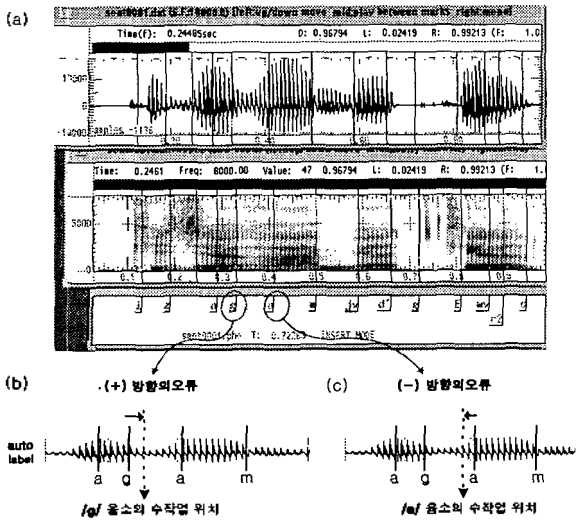

그림 1. (a) 자동 레이블링 오류 예 : "기자가 몇 개월",
(b), (c) : 음운 환경에 따른 오류의 방향성,
(b) '모음+자음', (c) '자음+모음'

Fig.1. (a) An example of an error by automatic segmentation,
(b), (c) : The biased error caused by phoneme environments,
(b) The case of 'vowel+consonant', (e.g., /a/_/g/),
(c) The case of 'consonant+vowel', (e.g., /g/_/a/).

III. 제안된 MLP 기반 후처리 시스템

본 절에서는 자동 분절 경계의 오류 범위를 줄일 수 있는 후처리에 대해 제안한다. 즉, 수작업으로 보정된 데이터의 특징 벡터를 다층 신경회로망을 통해 학습을 한 후, 자동 분절 결과와 후처리기인 MLP를 이용하여 새로운 음소 경계를 추출하는 일련의 과정에 대해 기술한다.

3.1. 음성 데이터베이스

본 논문에서는 한국 전자통신 연구원에서 제공한 문장 단위 음성 합성용 운율 데이터 베이스를 사용하였다. 남성 화자 1인이 낭독체로 발성하였으며, 녹음된 음성 데이터는 16kHz, 16바트의 해상도를 갖는 디지털 신호로 A/D 변환하였다. 음성 데이터는 우선 자동 레이블링 시스템을 이용하여 모든 음성 데이터를 음소 경계로 분할하고, 다음에 자동으로 분절된 음소의 경계를 수작업으로 보정하였다. 여기서, 신경망의 훈련 데이터로 쓰기 위해 100문장만이 수작업에 의해 보정된다.

본 실험에 사용된 훈련 데이터는 총 2,092문장에서 90문장(7,426 phonemes)을, 성능 평가를 위한 테스트 데이터는 100문장(8,204 phonemes)을 사용하였다.

3. 2. 특징벡터 추출[9][10][11][12][13]

음소분할에서 가장 중요하게 고려해야 할 사항은 음향적 변이 특성을 비교적 잘 표현하는 스펙트럼 변화 특성을 이용하되 음소간 변화에 민감하면서 화자 특성에는 둔감하도록 선정되어야 한다. 또한 음소간 변별력이 뛰어나면서도 음성학적으로 중요하지 않은 변화 요인에는 둔감한 특징을 가지는 특징 파라미터의 선정이 요구된다. 아울러 조음방법 및 조음위치에 따른 음소 고유의 특징 뿐아니라, 전후 음소에 의한 음소정보의 중첩 등이 고려되어야 한다. 또한, 이러한 음향학적 정보가 경계에서 뚜렷이 나타난다는 사실을 감안할 때 각 특징들의 근접 프레임과의 차이 정보가 중요하며, 음소와 음소간 천이는 시간적으로 서서히 발생하는 음운 환경을 고려할 때 많은 좌우 프레임 정보를 포함할수록 경계 결정에 도움을 준다.

이에 근거해 시간영역에서 나타나는 파형 정보인 영교차율, 에너지, 주파수 영역에서 나타나는 스펙트럼 정보인 대역별 에너지 정보 등 9차를 특징벡터 I으로 하고, 사람의 청각 특성을 모델링한 PLP 계수 13차를 특징벡터 II으로 하여 음성 특징을 추출하였다. 여기서, 특징벡터 I의 경우 특징 파라미터들 간의 변화량을 의미하는 근접 프레임 특징과의 절대차를 구하였으며, 통계 특성을 이용하여 정규화 되었다.

표 1. 음성 분석 조건 및 특징 파라미터

Table 1. Speech analysis conditions and features.

분석 조건	Samplingrate	16 [KHz]
	A/D quantization	16 [bits]
	Window type	Hamming window
	Window size	16 [msec]
	Overlap interval	6 [msec]
특징벡터 I	Energy	Frame log energy (1 order)
	Zero crossing rate	Frame zcr (1 order)
	A ratio low to high band energy	Low(0-3KHz), high(3-7KHz) (1 order)
	Band energy	Log band energy of 0-7[KHz] per 1[KHz] (6 order)r
특징벡터 II	PLP	13 order

3. 3. 특징벡터의 MLP 훈련[8][9][14]

3.3.1. 다층 신경 회로망 구조

3.2절에서 추출된 음소 경계간 음성학적 특징을 MLP를 이용하여 학습하기 위해, 본 실험에서는 입력층, 1개의 은닉층, 출력층으로 구성된 다층 신경 회로망을 채택하였다.

입력층은 경계 주변 영향(context effect)을 고려하기 위해, 분석 프레임의 좌측 한 프레임과 우측 두 프레임, 즉, 4 프레임(40msec)의 특징 벡터가 하나의 입력 패턴이 된다. 이는 유사한 음소들간 천이가 경계 구분이 불분명할 정도로 서서히 진행되므로 좌우 프레임의 특징을 많이 포함할수록 경계추출에 용이하기 때문이다. 특징벡터 I의 경우는 총 4(frame) x 9(order) + 1개의, 특징벡터 II의 경우 4(frame) x 13(order) + 1개의 입력 노드를 가지

게 된다. 은닉층은 실험에 의해서 결정된 15개의 노드로 구성되고, 출력층은 경계 유/무를 결정하는 1개의 노드를 가진다.

3.3.2. 특징벡터 I, II의 훈련

훈련 데이터는 수동분절 데이터의 음소 경계 유무에 따라 출력노드에 각각 1 또는 -1을 할당하고, 경계 주변 영향을 고려하기 위해 분석 프레임을 제외한 좌우 프레임에 경계가 존재할 경우 -0.01을 할당하여 작성한다. 그림 2는 특징 벡터 I, II에 대한 훈련과 최적의 가중치(weighting) 추출 과정을 블록 다이어그램으로 나타내었다. 우선 수동 분할 데이터를 이용하여 자동 레이블링 시스템의 오류 특성, 즉, 오류의 평균, 표준편차를 추출하고 각 특징벡터에 대해 MLP훈련한다.

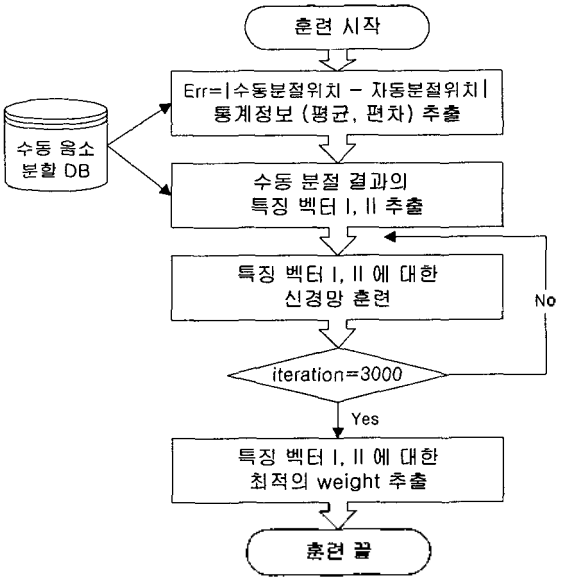

그림 2. 최적의 가중치 추출을 위한 MLP 훈련 과정
Fig. 2. The MLP training procedure for optimal weight extraction.

3.3.3. 특징벡터 I, II의 MLP 분절 성능

표 2는 기존 시스템인 음성언어 번역 시스템과 특징벡터 I, II의 MLP훈련 후 분할성능을 음운 환경별로 비교하였을 때, 상대적으로 우세한 성능을 보이는 특징벡터별 음운환경을 제시하였다. 여기서, 우세한 성능이란 각각 1 frame accuracy가 60%이상인 경우를 의미하며, 1 frame accuracy는 수동분절 위치로부터 분절결과의 오류 범위가 ±1 frame(±15 msec)내로 분할될 때를 말한다. 표 2의 결과를 정리하면 다음과 같다.

자동 분할기의 성능 및 특징벡터 I, II을 이용한 MLP 분절 성능이 음소 또는 음운환경에 따른 성능 차이를 보인다. 다시 말해 자동 분할기는 다른 특징벡터에 비해 비음부의 분절 성능이 우수하고, 특징벡터 I은 무성 자음부에, 특징벡터 II는 '모음+모음' 을 제외한 모음부에서 성능이 우수했다. 이는 주로 시간영역에서 추출된 특징 파라미터들로 이뤄진 특징벡터 I의 경우, 대체로 유/무성 구분과 자/모음 구분에 두드러진 특징을 나타내었으며, 특징벡터 II의 경우, 모음의 제 1포만트, 제 2포만트를 잘 표현하는 PLP 분석으로[13], 모음부와 '유음+유음' 에서 우수한 분할 성능을 보였다. 이로써, 특정 음운 환경에 강인한 feature를 도입하여 분할한다면 합성 DB의 자동 구축에 기여할 수 있을 것이라 기대된다.

표 2. 각 특징벡터에 우수한 음운환경
Table 2. Phoneme groups excellent at each feature vector.

	특징 벡터 I	특징벡터 II	자동분절기
음 운 환 경	[유음,무성 자음] +모음	[비음,유음,무음]+모음 비음+비음	비음부
		유음+유음	모음+유음
	무성 자음부	비음+무성 자음	

3. 4. 제안된 음소 경계 추출 과정

후처리는 테스트 데이터의 자동 음소 분절 경계를 MLP 출력값을 이용하여 새로운 음소 경계를 결정하는 것을 말한다. 즉, 자동 분절 경계로부터 일정 범위를 탐색구간으로 하여, 임계값 이상인 MLP 출력값 중 최대값을 가지는 위치를 후처리된 경계로 선택한다. 제안된 후처리 알고리듬에 대한 구체적인 방법은 다음과 같다 (그림 3).

Step 1. 해당 음소의 음운 환경별 특징벡터를 선정한다.
Step 2. 자동 분절 경계 위치로부터 좌우 2 frame (± 25 msec)을 MLP 출력값 탐색구간으로 결정한다.
Step 3. 탐색구간 내에 임계값 이상인 MLP 출력값 중 최대값 두개를 결정한 후 자동 분절 위치에 더 가까운 위치를 후처리된 경계로 본다.
Step 4. step 3과정이 실패할 경우, 음운환경에 따른 통계 자료가 등록된 테이블로부터, 90%의 신뢰 구간까지 탐색구간을 확장한다.확장된 탐색구간에서 step 3과정을 반복한다.
탐색구간의 확장 범위는 아래 식에 따른다.
Unit: [msec]
If average of error<0
[*average-standard-eviation* × *1.645*,*25*] (1)
Else
[-25, average+standard deviation × 1.645] (2)
Step 5. step 4과정이 실패할 경우, 자동 분절 결과를 음소 경계 위치로 한다.

각 과정에 대해 구체적으로 기술하면 다음과 같다.

우선, step 1에서 음운환경별 특징벡터는 표 3에서 제시된 type에 의해서 선정된다. 이는 3.3.3절에서 특징 벡터 I, II 및 자동 음소 분절기의 MLP분절 성능을 분석한 결과, 가장 우수한 성능을 보이는 음운 환경을 의미한다. 따라서, 해당 음소의 음운 환경에 따라 특징 벡터를 선정한다.

step 4는 기준이 되는 자동 분절 결과가 큰 오류 범위를 가지기 때문에, 임의로 설정된 탐색구간 2 frame(±25msec)의 확장이 요구되며 구체적인 방법은 다음과 같다. 자동 분할 시스템의 오류 특성 (2절 참고)을 후처리의 탐색구간 확장에 도입하기 위해, 우선 수동 분절 결과와 자동 분절 결과의 통계적 특성을 분석한다. 이로부터 음운 환경에 따른 오류의 방향성, 평균 오류 및 오류의 표준편차가 등록된 테이블을 이용하여 탐색구간을 확장한다. 이때, 음소 경계의 오류분포를 정규분포라 가정하고 90% 신뢰구간까지 확장하며 그 범위는 식 (1) 또는 (2)을 이용한다.

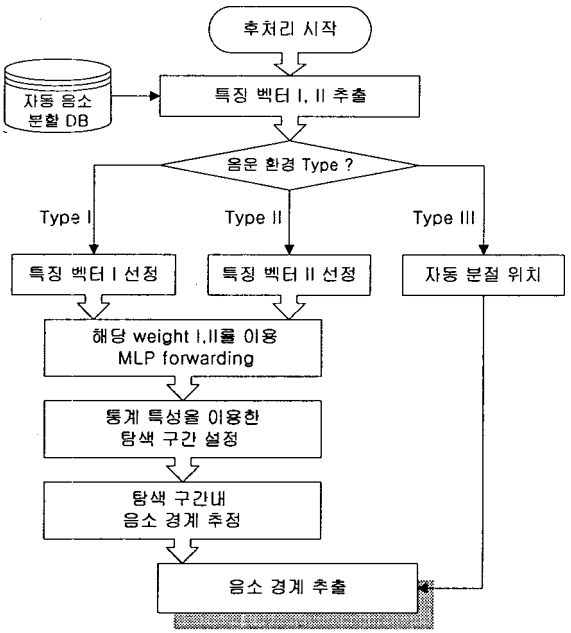

그림 3. 음소 경계 결정을 위한 후처리기
Fig. 3. The postprocessor for the phoneme boundary decision.

표 3. 음운 환경별 type 정의
Table 3. Definition for phoneme group type.

	Type I	Type II	Type III
음운 환경	무 성 자음부	[비음,유음, 무성음]+모음	모음+모음
		비음+비음	[모음,유음,무성음]+비음
		유음+유음	[모음,비음,무성음]+유음

IV. 실험 및 결과

제안된 자동 음소 분할기의 후처리기를 평가하기 위해, 문장단위 합성 운율 데이터 베이스에 제안된 방법을 적용한 후 프레임 분할 성능 및 절대 오류를 각각 측정, 비교하였다. 또한 자동 분절결과를 그대로 이용한 경우와 제안된 방식을 이용한 합성음의 주관적인 음질 평가를 실시하였다.

4. 1. 낭독체 문장에 대한 후처리 성능

표 4는 낭독체 문장에 대해 후처리 이전과 특징 벡터 I, II의 후처리 이후 성능에 대한 비교 분석이다. 특히, 해당 음운 환경에 강한 특징벡터를 이용한 후처리의 성능을 비교하였다. 즉, 음운 환경별 MLP 분절 성능을 미리 분석하여, 자동 분할기가 가장 우수한 성능을 보이는 음운환경에 대해서는 자동 분절 결과를 그대로 사용하고, 그 외의 경우 해당 음운 환경에 강인한 특징벡터를 적용하여 후처리 과정을 수행한 결과를 말한다(그림 3 참조). 특징 벡터 I, II에 대해 단독으로 후처리를 적용한 결과보다 음운환경에 따라 특징벡터의 선택이 이뤄진 결합된 형태의 후처리 성능이 우수하였다. 후처리 이전인 자동 분할기 성능에 비해 49.6%에서 69.5%로 약 19.9%의 프레임 분할률 향상을 보였으며, 절대오류는 17.9 msec에서 12.9 msec로 약 28.6% 감소되었다. 여기서 on frame accuracy, 1 frame accuracy 및 2 frame accuracy는 수작업된 경계로부터 오류 범위가 각각 ±5ms이내, ±15ms이내, ±25ms이내로 분할됨을 의미한다.

이는 제안된 후처리기로 자동 음소 분할 오류의 범위를 줄일 수 있음을 보이는 것이다. 비록 고립단어에 비해 떨어지는 성능을 얻었지만[3], 합성단위 추출 자동화에 기여할 수 있을 것이다. 연속체가 고립단어에 비해 성능 차이를 보이는 것은 다양한 음운 환경이 존재하는 문장인 경우 상대적으로 훈련되는 각 음운환경의 개수가 적으므로 MLP 분할 성능이 전반적으로 떨어지는데 기인한 것이라 본다. 그림 4는 자동 분할의 예와 후처리 결과를 나타내었으며, 자동 분절 결과가 후처리에 의해서 이동한 경우 '+' 기호로 구분하였다.

표 4. 프레임 분할 성능 및 절대 오류
Table 4. The performance of frame segmentation accuracy and absolute error.

Postprocessing		Before	After		
			특징벡터 I	특징벡터 II	자동분할기+ 특징벡터 I, II
분할률 [%]	On frame	26.82 [2,200/ 8,204]	33.40 [2,740/ 8,204]	32.23 [2,644/ 8,204]	35.52 [2,914/8,204]
	±1 frame	49.62 [4,071/ 8,204]	66.22 [5,433/ 8,204]	65.71 [5,391/ 8,204]	69.51 [5,703/8,204]
	±2 frame	67.09 [5,504/ 8,204]	78.97 [6,479/ 8,204]	79.03 [6,484/ 8,204]	81.69 [6,702/8,204]
절대오류[msec]		17.91	13.37	13.63	12.36

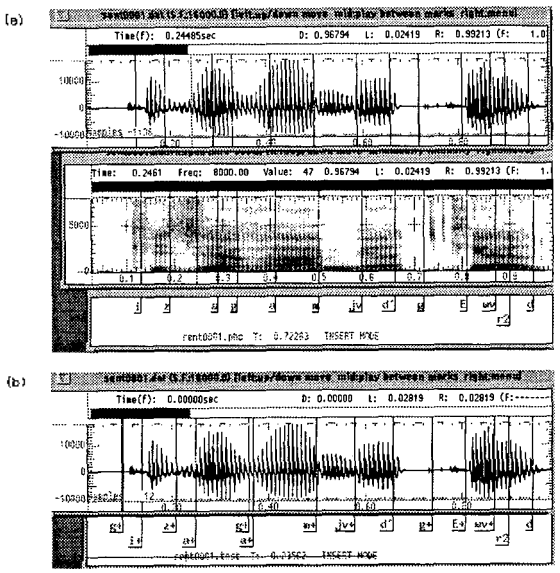

그림 4. 자동 분할 오류의 예와 후처리후 분할결과, "기자가 몇 개월",
(a) 자동 분할의 예, (b) 후처리 후 분할 결과
Fig. 4. (a) Example of error by auto label, (b) The result of new label.

4. 2. 청취 테스트에 의한 주관적인 음질평가

기존 시스템의 결과와 제안된 방법의 결과를 비교하기 위해 청취 테스트를 이용하여 합성음 평가를 실시하였다. 합성음의 음질은 명료도와 자연성 두 가지가 모두 고려되어야 한다. 명료도란 청취자가 음성에 포함된 단어를 얼마나 잘 알아들을 수 있는가이고, 자연성이란 음의 높낮이, 지속시간과 매끄러운 정도 등 쉽게 정의 되기 힘든 막연한 개념을 말한다.

따라서, 본 실험에서는 기존 시스템과 제안된 방법을 이용한 각각의 합성음을 문장 단위로 제시하여 명료도와 자연성을 평가하였다.

4.2.1. 합성음 생성 및 청취 테스트

청취 테스트는 임의의 문장(표 5)으로 생성된 합성음을 비전문가 10명에게 블라인드 테스트 (blind test)형태로 실시되었으며, 명료도와 자연성 항목에 대해 5점에서 1점까지 평가하도록 하였다. 합성음은 다음 (A), (B) 방법으로 접합점에서의 에너지 정규화 및 smoothing 등 동일 조건에서 생성된다.

(A) 자동 분절된 데이터를 그대로 이용할 경우
(B) 자동 분절된 데이터에 제안된 방법을 적용한 경우

4.2.2. 청취 테스트 결과

합성음 생성 방법 (A), (B)에 대한 명료도/자연성의 항목별 주관적 평가 결과는 표 6에 나타내었다. 평가 결과에서 보았듯이, 명료도에는 (A), (B) 방법 모두가 양호한 음질(약 3.4)을 얻었으며, 자연성 평가에서는 평균적으로 (B)가 우수한 성능을 보였다. 그러나 표준 편차에서 알 수 있듯이, 제안된 방법 (B)는 문장에 따라 상대적으로 많은 성능 차이를 보였다. 이는 자동 분절기의 일관된 분절 오류를 인위적인 후처리에 의해 보정함으로써 발생하는 문제임을 알 수 있었다.

표 5. 합성음 평가에 사용된 문장
Table 5. Sentences used for the listening tests.

Nun	합성 문장
1	한시간 지나서가 문제입니다
2	앞으로 어찌할 것인가가 걱정이다
3	여기서부터가 서울입니다
4	할아버지께서 이름을 지어 주셨다
5	이번에도 우리 학교에서 우승을 했다.
6	3학년 학생들이 도서관 앞 청소를 한다
7	청소는 우리반이 가장 깨끗해요
8	들이서 그 일을 다 마무리 했다
9	지구는 태양의 둘레를 돌고 있다
10	할아버지께서는 시조를 잘 읊으셨습니다

표 6. 합성음 청취 테스트 결과
Table 6. The listening test results.

		1	2	3	4	5	6	7	8	9	10	평균	표준편차
명료도	(A)	3.6	3.2	3.4	3.3	3.5	3.1	3.5	3.7	3.9	3.0	3.40	0.29
	(B)	3.5	3.1	3.9	4.1	4.0	3.4	2.8	3.0	2.9	3.1	3.37	0.49
자연성	(A)	2.9	2.8	3.1	2.7	3.3	2.7	2.9	3.0	3.2	3.6	3.03	0.29
	(B)	3.5	2.9	4.3	3.9	3.4	2.8	3.1	2.9	3.1	3.3	3.33	0.50

V. 결 론

본 논문은 일정 범위의 오류 및 일관성 있는 오류를 포함하는 기존 음성 번역 시스템의 자동 분절된 결과를 MLP 학습을 통해 경계 오류를 보상할 목적으로 제안되었다. 자동 분할기가 수작업보다 일관성을 유지하는 장점을 가지나 큰 오류를 포함하여 합성단위로 직접 사용되기에 부적합하다. 따라서, MLP 기반 후처리기로 음소경계 위치를 세밀하게 이동함으로써 오류의 범위를 줄일 수 있었으며, 이는 본 논문의 궁극적인 목표인 합성단위 자동 생성에 기여에 대한 가능성을 제시하였다. 제안된 후처리를 도입한 결과, 자동 분할기의 성능에 비해 약 19.9%의 frame accuracy 향상과 28.6%의 절대 오류 향상을 보였다.

그러나, 제안된 방식에 여전히 포함되는 몇 가지 오류가 존재한다.

첫째, MLP 탐색 구간은 자동 분절 결과를 기준으로 설정되는데, 이때 큰 오류를 포함한 자동 분절 경계가 기준인 경우 후처리로 보정할 수 있는 범위를 넘어선다.

둘째, 탐색 구간 확장을 위해 통계 자료가 등록된 테이블

로부터 음운환경에 대한 정보를 가져오는데, 이때, 자동 분할기의 내재된 grapheme-to-phoneme 오류로 인해 잘못된 음운환경 정보를 가져오면서 발생하는 오류를 포함하게 된다.

셋째, 탐색구간의 설정은 자동 분할기의 오류가 정규 분포라 가정하고 수행하였다. 그러나, 대체로 분산이 큰 경우, 즉, 빈도수가 적은 음운환경 및 오류가 불규칙한 방향으로 발생하는 경우에 대한 보다 체계적인 고려가 필수적이다.

넷째, 자동분절 결과를 그대로 이용한 합성음 음질은 비슷한 성능(약 3.4)을 유지했으나, 제안된 방식이 적용된 합성음질은 문장에 따라 다소 큰 성능차이를 보였다. 이는 자동 분할기에 비해 경계오류는 수정되었지만, 자동 분할기의 가장 큰 장점인 일관성 유지가 어려웠기 때문이다.

이를 개선하기 위해, 향후 후처리에 의한 경계 보정을 하거나, 에너지가 최소로 되는 접합점을 찾는 방법 등 다양한 측면에서 연구가 진행 되어야 한다. 또한, 수동 레이블링 작업에서 얻은 지식과 경험으로 정확한 자동 레이블링 시스템의 종합적인 분석과 더불어, 다양한 음운 환경에 대한 구체적인 그룹화 및 충분한 분석이 경계 결정에 이용된다면 자동 분할기의 성능 향상에 기여할 것이다.

본 논문에서 제안된 방식은 수동 분할에 비해 시간과 노력을 상당히 줄일 수 있으며, 음소 경계 위치의 수정이 간단하여 대량의 한국어 음성 데이터 베이스 구축에 기여할 것이다. 또한, 합성 데이터베이스 제작의 소요 시간 단축으로 새로운 화자에 대한 합성기 구현이 용이할 것이다.

참 고 문 헌

1. Nick Campbell, "Processing a Speech Corpus for CHATR Synthesis," *Pro. ICSP*, pp.183-186, 1997.
2. 김상훈, 이정철, 강동규, 이영직, "대용량 운율 음성 데이터를 이용한 자동합성방식," 제15회 음성통신 및 신호처리 워크샵, pp.87-92, 1998.
3. 박은영, 김상훈, 정재호, "합성 단위 자동 생성을 위한 자동 음소 분할기 후처리에 대한 연구," 한국음향학회지, 제17권, 제7호, pp. 50-56, 1998.
4. A. W. Black & Nick Campbell, "Optimizing Selection of Units from Speech Databases for concatenative Synthesis," *EUROSPEECH*, pp.581-584, 1995.
5. Nakajima S. and Hamada H., "Automatic generation of synthesis unit based on context oriented clustering," *Proc. ICASSP*, pp.659-662, 1988.
6. T. Svendsen and Frank K. Soong, "On the Automatic Segmentation of Speech Signals," *Proc. ICASSP*, pp.77-80, 1987.
7. 김상훈, 이창섭, 김희린, "운율 분석용 DB작성을 위한 자동 레이블러의 성능 평가 및 유용성," *SICOPS96* SESSON 3.6, 1996.
8. J.P. Marten and L. Depuydt, "Broad phonetic classification and segmentation of continuous speech by means of neural networks and dynamic programming," *Speech ommunication*, pp.81-90, 1991.
9. Y. Suh and Y. Lee, "Phoneme Segmentation of Continuous Speech Using The Multi-Layer Perceptrons," *Proc. ICSLP*, pp.1293-1296, 1996.
10. Victor W. Zue, "The Use of speech knowledge in automatic speech recognition," *Proceeding of the IEEE*, pp. 1602-1615, 1985.
11. Ronald A. Cole and Llly Hou, "Segmentation and Broad Classification of Continuous Speech," *Proc. ICASSP*, pp.453-456, 1988.
12. David B.Grayden and Michael S. Scordilis, "Phonemic Segmentation of Fluent Speech," *Proc. ICASSP*, pp.73-76, 1994.
13. H. Hermansky, B. A. Hanson and H. Walkita, "Perceptually Based Linear Predictive Analysis of Speech," *Proc. ICASSP*, pp. 509-512, March 1985.
14. Richard P. Lippmann, "An Introduction to Computing with Neural Nets," *IEEE ASSP magazine*, pp.4-22, April 1987

▲박 은 영(Eun-Young Park) 1973년 9월 29일생

1996년 2월 : 인하대학교 전자공학과 학사
1999년 2월 : 인하대학교 전자공학과 석사
현재 : LG 정보통신 중앙연구소, 연구원
※ 주관심분야 : 음성합성, 음성인식

▲김 상 훈(Sang-hun Kim) 1967년 10월 1일생
1990년 2월 : 연세대학교 전기공학과 학사
1992년 2월 : KAIST 전기 및 전자공학과 석사
현재 : 한국전자통신연구원 음성신호처리연구실 선임연구원
※ 주관심분야 : 음성합성, 음성인식

▲정 재 호(Jae-Ho Chung)
1982년 : 미국 University of Maryland (학사)
1984년 : 미국 University of Maryland (석사)
1990년 : 미국 Georgia Institute of Technology(박사)
1984년~1985년 : 미국 국방성 산하 해군 연구소, 신호처리실 연구원
1991년~1992년 : 미국 AT&T Bell Labolatories, 음성 신호처리 연구실, 연구원(MTS)
1992년~현재 : 인하대학교 공과대학 전자공학과, 현(부교수)