합성단위 자동생성을 위한 자동 음소 분할기 후처리에 대한 연구

The Postprocessor of Automatic Segmentation for Synthesis Unit Generation

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

요 약

본 논문은 자동 음소 분할기의 음소 경계 오류를 보상하기 위한 후처리(Postprocessing)에 관한 연구이다. 이는 현재 음성 합성을 위한 음성/언어학적 연구, 운율 모델링, 합성단위 자동 생성 연구 등에 대량의 음소 단위 분절과 음소 레이블링된 데이터의 필요성에 따른 연구의 일환이다. 특히 수작업에 의한 분절 및 레이블링은 일관성의 유지가 어렵고 긴 시간이 소요되므로 자동 분절 기술이 더욱 중요시 되고 있다. 따라서, 본 논문은 자동 분절 경계의 오류 범위를 줄일 수 있는 후처리기를 제안하여 자동 분절 결과를 직접 합성 단위로 사용할 수 있고 대량의 합성용 운율 데이터 베이스 구축에 유용함을 기술한다.

제안된 후처리기는 수작업으로 보정된 데이터의 특징 벡터를 다층 신경회로망 (MLP; Multi-layer perceptron)을 통해 학습을 한 후, ETRI(Electronics and Telecommunication Research Institute)에서 개발된 음성 언어 번역 시스템을 이용한 자동 분절 결과와 후처리기인 MLP를 이용하여 새로운 음소 경계를 추출한다.

고립단어로 발성된 합성 데이터베이스에서 후처리기로 보정된 분절 결과는 음성 언어 번역 시스템의 분할율보다 약 25%의 향상된 성능을 보였으며, 절대 오류(|*Hand label position-Auto label position*|)는 약 39%가 향상되었다. 이는 MLP를 이용한 후처리로 자동 분절 오류의 범위를 줄일 수 있고, 대량의 합성용 운율 데이터 베이스 구축 및 합성 단위의 자동 생성에 이용될 수 있음을 보이는 것이다.

ABSTRACT

The work presented in this paper is about a postprocessor, which improves the performance of automatic segmentation system when phoneme boundary errors exist. Segmented and labeled speech data are essential for phonetic research as well as for speech synthesis system construction based on the segmental concatenation. However, manual segmentation and labeling are not efficient due to the lack of consistency, and also they are time consuming. For that reasons, we propose a postprocessor which reduces the range of errors in the auto labeled results which are then going to be used directly as synthesis unit. Starting from a baseline automatic segmentation system, 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.

We have achieved 25% improvement for the frame accuracy, comparing with the performance of automatic labeling system for isolated speech. Also, we could reduce the absolute error rate about 39%. These results show that we can compensate boundary error rate by using proposed MLP postprocessor, and expect contribution to constructing various prosodic synthesis DB and to generating automatic synthesis unit in speech synthesis area.

I. 서 론

자동 음소 분할(Automatic phoneme segmentation)은 음소단위 음성인식이나 이를 위한 음소 분할된 훈련 데이터베이스 구축에 이용되는 기술로 음성합성 분야에서도 대용량 합성용 운율 데이터베이스 구축에 이용되고 있다. 수동 분할에 비해 능률적인 작업이 이뤄질 수 있고, 일관성 있는 음소 단위 분절 결과를 줄 수 있어 더욱 중요해진 자동 분절 기술은 다양하게 연구 되어져 왔다. 대표적인 방식으로 패턴 매칭 방법, 스펙트럼 변화 특성을 이용하는 방법과 constrained clustering vector quantization 방법 등이 있다[1]. 이러한 꾸준한 연구에 따른 성능 향상으로 음성합성 분야에서는 자동 음소 분절 결과를 직접 합성 단위로 사용하고자 하는 연구가 시도되고 있으나, 자동 분절의 음소 경계 오류로 인한 접합 경계에서의 불연속성

* 인하대학교 전자공학과
** 한국전자통신연구원
접수일자 : 1998년 7월 4일

등의 왜곡을 고려해야 하는 문제점을 가지고 있다[2]. 따라서 합성 단위 자동 생성을 위해서는 음소분할기의 성능을 향상시키거나 합성단위 연결시 최적 접합점을 찾는 기술이 필요하다[2][3][4]. 또한, 최근 합성연구의 주류가 되고 있는 대용량 합성 데이터 베이스로부터 가장 적절한 합성 단위간 접합 방식으로 인해 인식기를 이용한 자동 분절 기술의 중요도가 더욱 증대되고 있으며, 현재 운율 현상이 고려된 합성 단위를 이용하여 합성음의 자연성을 유지하기 위한 합성기에 대한 연구가 진행되고 있다[5].

음성 언어 번역 시스템을 이용한 자동 음소분할기의 성능은 문장단위의 발화에서 약 80%가 절대 오류 30 msec 이내이며, 고립 단어인 경우 그 성능은 더욱 떨어진다[3]. 따라서 본 논문은 기존 음소분할기의 성능을 향상시키기 위한 연구로, 자동 분절 결과를 다층 신경회로망을 이용하여 후처리 함으로써 경계 오류의 범위를 줄이고, 궁극적으로 대량의 합성용 운율 데이터 베이스 구축 및 합성 단위의 자동 생성에 사용하고자 한다. 특히 자동 분할된 음소는 그 음소의 안정구간이 항상 포함되기 때문에 국부적인 두 음소간 음성학적 지식(Phonetic knowledge)을 최대한 이용하면 음소경계 오류를 최소화할 수 있다[6]. 그러나 매우 다양하게 나타나는 음소 경계간 음성학적 특징을 단시간 내에 적용할 수는 없기 때문에, 수동 분절된 결과를 MLP 훈련하여 음소경계를 보정한다[7].

본 논문의 구성은 음소의 음향적 특징을 추출하는 전처리와 다층 신경회로망으로 학습하여 음소 경계 유무를 출력하는 훈련단계 및 자동 분절 결과와 다층 신경회로망 출력값을 이용하여 정확한 음소 경계를 결정하는 후처리 과정으로 이루어진다. 그림 1은 제안된 후처리가 포함된 음성 분할 시스템의 전체 구성도이다.

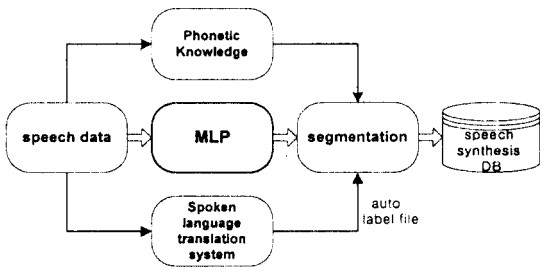

그림 1. 자동 분절기의 후처리
Fig. 1 Postprocessor of the automatic segmentation system.

II. 특징 파라메터 추출

음성학적 측면에서 음소의 특징을 구분하는 기준은 일반적으로 시간영역에서 나타나는 파형정보인 영교차율, 에너지, 자기 상관 계수, 피치 등이 있고, 주파수 영역에서 나타나는 스펙트럼 정보인 대역 에너지, 포만트 주파수 뿐 아니라 조음 방법, 조음 위치에 따라 변하는 스펙트럼 천이 정보 등이 있다[8]. 합성을 위한 음소분할에서 가장 중요하게 고려해야 할 사항은 음향적 변이 특성을 비교적 잘 표현하는 스펙트럼 변화 특성을 이용하되 음소간 변화에 민감하면서 화자 특성에는 둔감하도록 선정되어야 하고 음소간 변별력이 뛰어나면서도 음성학적으로 중요하지 않은 변화 요인에는 둔감한 특징을 가지는 특징 파라메터의 선정이 요구된다[6]. 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[7][9][10].

본 논문의 실험에서 사용된 음성 데이터는 남성 화자가 발성한 고립 단어 합성용 데이터베이스이며, 음성 분석 및 특징 파라메터 추출조건은 표 1에 나타내었다[7][11][12]. 표 1의 특징 파라메터는 통계 특성을 이용하여 정규화 되었다.

표 1. 음성 분석 조건 및 특징 파라메터
Table 1. Speech analysis conditions and features.

Analysis condition	Sampling rate	16 [kHz]
	A/D quantization	16 [bits]
	Window type	Hamming window
	Window size	16 [msec]
	Overlap interval	6 [msec]
Features	Energy	Frame log energy (1 order)
	Zero crossing rate	Frame zcr (1 order)
	A ratio low to high band energy	Low(0-3000Hz), high(3000-7500Hz)(1 order)
	Band energy	Log band energy of 0-7500[Hz] per 1[kHz] (6 order)

III. 다층 신경회로망 훈련

다층 신경망은 입력층, 1개의 은닉층, 출력층으로 구성되며 입력층에는 음소간 천이 구간을 고려하기 위해 3프레임(30msec)의 특징 벡터가 하나의 입력 패턴이 된다. 은닉층은 13개의 은닉 노드로 구성되며 출력층은 음소 경계의 유무를 결정하는 1개의 노드로 구성된다. 3프레임 중 가운데 프레임의 음소 경계 유무에 따라 출력 노드 값을 결정한다. 즉, 훈련 데이터는 음소 경계의 유무에 따라 출력노드에 각각 0.9 또는 −0.1를 할당하고, 경계 주변 영향을 고려하기 위해 가운데 프레임을 제외한 좌우 프레임에 경계가 존재할 경우 0.3을 할당하여 작성한다.

IV. 자동 분할기의 후처리

후처리는 테스트 데이터의 자동 음소 분절 경계를 MLP 출력값을 이용하여 새로운 음소 경계를 결정하는 것을 말한다. 즉, 임계값 이상인 MLP 출력값 중 최대값을 가지는 위치를 후처리된 경계로 선택한다. 이때, MLP 탐색 구간은 자동 음소 분절 경계로부터 일정 범위로 결정된다.

4.1 음운환경에 따른 음소 경계의 오류

자동 분할기의 성능을 음소별, 음운환경별로 볼 때, 음소별로는 모음보다는 자음의 레이블링 정확률이 높게 나타나고, 비음부에서 비교적 정확하게 레이블링 된다. 그러나 음운환경 '모음+모음', '모음+유음', '파열/파찰음+모음'의 경우 음소 경계에서 오류가 크게 나타난다[3]. 본 논문에서 사용한 음소 레이블링 기호를 표 2에 나타내었다. 자동 분할기의 성능은 음소 또는 음운환경에 따른 성능 차이를 보이며 표 3, 4에서 보이는 것과 같이 음소 경계의 오류가 일정한 방향성을 가지면서 분절되는 것을 알 수 있다. 이는 자동 분할기는 수작업에 의한 분절에 비해 일관성을 유지한다는 것과 의미를 같이한다[1]. 따라서 이렇게 한쪽 방향으로 치우친 자동 분절 결과를 그대로 이용한 합성음은 경계에서의 불연속(segment discontinuity or discontinuity distortion)으로 인한 스펙트럼 왜곡이 발생하여 합성음의 음질이 떨어진다[2]. 즉, 자동 분절 결과의 일관성을 최대한 유지하되 경계에서의 오류를 보정한다면 합성 단위간 연결점에서의 왜곡은 최소화 될 수 있다.

본 논문에서 채택한 MLP 탐색 구간은 자동 분절 경계를 기준으로 설정되는데, 표 3, 4에서 보이는 경계 오류방향의 일관성을 이용한다면 더욱 향상된 성능을 기대할 수 있을 것이다.

4.1.1 '(모음, 종성 비음(/N/, /m2/, /n2/), 종성 유음(/r2/))+자음'의 음운 환경.

무성 자음이 모음, 종성 비음, 종성 유음 다음에 위치하게 되면 유성음화 되기 쉽다. 이로 인해 자음의 특성이 앞의 음운인 유성음의 영향으로 해당 음소 고유의 음향적 특징이 상쇄되어 경계에서 뚜렷한 음향적인 정보가 나타나지 않는다. 즉, 주어진 음소의 음향적 특징이 전후 음소에 따라 변하는 동시조음효과(coarticulation or contextual influence)로 음소 정보(phonetic information)가 중첩되어 경계에서 그 특징이 상쇄되는 것이다[6].

따라서, 음향적 변화를 기준으로 레이블링 되는 자동 분할기는 음소 경계에서의 차이 또는 천이 정보를 정확히 검출하지 못한다. 그 예로 표 3에서 언급된 무성 자음이 모음, 종성 비음, 종성 유음 다음에 위치할 때, 자음 경계의 80%이상이 유성음부 쪽으로 치우쳐서 분절되어 오류의 방향이 양이 되는 경우를 볼 수 있다. 즉, 자동 분절 위치가 수동 분절 위치에 비해 왼쪽에 위치하게 되는 것을 말한다. 이렇게, 유성음이 포함된 무성음을 직접 접합할 경우 합성음의 음질은 저하를 가져온다.

4.1.2 '자음+모음'의 음운 환경

표 4에서 언급된 자음과 모음 환경에서, 모음의 경계의 80%이상이 원래 모음위치보다 더 앞쪽에 분절되어 음의 오류의 방향을 가진다. 다시 말해, 모음의 위치가 모음부로 치우치므로 자동 분절 위치가 수동 분절 위치에 비해 오른쪽에 위치한다.

표 2. 음소 기호 표기 테이블
Table 2. Table of phoneme symbols.

모음	표기법	자음	초성표기법	종성표기법
ㅏ	a	ㄱ	g	g'
ㅑ	ja	ㄲ	G	
ㅓ	v	ㄴ	n	n2
ㅕ	jv	ㄷ	d	d'
ㅗ	o	ㄸ	D	
ㅛ	jo	ㄹ	r	r2
ㅜ	u	ㅁ	m	m2
ㅠ	ju	ㅂ	b	b'
ㅡ	U	ㅃ	B	
ㅣ	i	ㅅ	s	
ㅚ, ㅞ	we	ㅆ	S	
ㅔ	jE	ㅇ		N
ㅘ	wa	ㅈ	z	
ㅟ	Wi	ㅉ	Z	
ㅐ	E	ㅊ	c	
ㅙ	e	ㅍ	p	
ㅒ	wE	ㅋ	k	
ㅢ	wi	ㅌ	t	
ㅖ	je	ㅎ	h	
ㅝ	wv			

표 3. '[모음, 비음, 유음]+자음'의 음운 환경에서 자동 분할기의 자음의 평균 오류
Table 3. Case of '[vowel, nasal (/N/, /m2/, /n2/) or liquid (/r2/)] +consonant: The average error of consonant in the automatic segmentation system (before postprocessing).

Phoneme	(+) direction of error [%]	(−) direction of error [%]	Average [msec]
r (liquid)	97.56	2.44	17.83
b (plosive)	97.85	2.15	25.51
d (plosive)	96.56	3.44	20.35
g (plosive)	96.50	3.50	22.04
s (fricative)	91.48	8.52	18.99
z (fricative)	91.85	8.15	12.95
D (plosive)	84.28	15.72	35.35
Z (fricative)	88.37	11.63	22.26
c (affricate)	86.19	13.81	11.87
p (plosive)	86.55	13.45	23.31
t (plosive)	86.55	13.45	23.31

표 4. '자음+모음'의 음운환경에서 자동 분할기의 모음의 평균 오류
Table 4. Case of 'consonant+vowel': The average error of vowel in the automatic segmentation system (before postprocessing).

Phoneme	(+) direction of error [%]	(−) direction of error [%]	Average [msec]
r (liquid)	5.90	94.10	29.81
Z (plosive)	6.52	93.48	−19.83

4.2 경계 결정을 위한 후처리 과정

자동 분절 경계를 기준으로부터 일정 범위를 MLP 탐색 구간으로 결정한다. 이때, 4.1절에서 언급된 자동 분절기의 오류의 통계 특성 이용하여 설정하게 된다. 즉, 본 실험에서 사용한 자동 분절기인, ETRI에서 개발된 음성언어 번역 시스템의 성능이 고립 단어의 경우 60%이상이 절대 오류 30 msec이내에 존재하므로 자동 분절 경계 위치로부터 좌우 2 frame (25 msec)을 MLP 출력값 탐색 구간으로 결정하였다[3]. 다음은 후처리를 이용한 음소경계 결정 과정을 나타낸다.

Step 1. 탐색구간(±25msec) 내에 임계값 이상인 MLP출력값 중 최대값의 위치를 후처리된 경계 위치로 한다.

Step 2. 탐색구간 내에 임계값 이상의 MLP 출력값이 존재하지 않을 경우, 음운환경에 따른 통계 자료가 등록된 표 5로부터, 90%의 신뢰 구간까지 탐색구간을 확장한다. 확장된 탐색구간에서 step 1과정을 반복한다.

탐색구간의 확장 범위는 아래 식에 따른다. Unit: [msec]

if Average of error <0

$$[Average-standard\ deviation*1.645,\ 25] \quad (1)$$

Else

$$[-25,\ Average+standard\ deviation*1.645] \quad (2)$$

Step 3. Step 2과정이 실패할 경우, 자동 분절 결과를 음소 경계 위치로 한다.

특히, step 2에서 기준이 되는 자동 분절 결과의 큰 오류 범위로 임의로 설정된 탐색구간(±25 msec)의 확장이 요구되며 구체적인 방법은 다음과 같다. 표 3, 4에서 보여준 것과 같이 음운환경에 따라 자동 분절 결과는 일정한 방향성을 가지고 분절되는 경향을 가진다. 자동 분할 시스템의 이러한 특성을 후처리의 탐색 구간 확장에 도입하기 위해, 우선 수동 분절 결과와 자동 분절 결과의 통계적 특성을 분석한다. 이로부터 음운 환경에 따른 오류의 방향성, 평균 오류 및 오류의 표준 편차가 등록된 테이블을 이용하여 탐색 구간을 확장한다. 이때, 음소 경계의 오류분포를 정규분포라 가정하고 90% 신뢰구간까지 확장하며 그 범위는 식 (1) 또는 (2)을 이용한다.

그림 2는 /r/_/o/ 음운환경에 대해 step 2과정을 보여준다. 음소 /o/의 자동 분절 결과를 기준으로 한 탐색 구간 내에서 step 1의 과정이 실패한 후, 등록된 통계 테이블 표 5로부터 음운환경 /r/_/o/에 대한 정보를 가져온다. 이때, 오류의 평균이 음의 방향성을 가지므로 식 (1)을 이용하여 음의 방향으로 탐색 구간 확장이 이뤄진다. 즉, 탐색 구간은 음소 /o/의 자동 분절 경계 위치로부터 음의 방향으로 7 frame까지 확장된다. 또한 그림 3은 자동 분절의 오류의 예와 MLP 출력값을 이용하여 후처리된 결과이며, 음운환경 /g/_/a/, /r/_/o/에 대해 음소 /a/, /o/의 경계는 후처리 후 각각 보정되었다.

표 5. 현재 음소의 평균 오류, 절대 오류, 표준 편차 및 신뢰구간 90% 범위의 자료가 등록된 통계 테이블 Unit: [msec]

Table 5. The registered statistics table, which is composed of average error of current phoneme boundary, absolute average error, standard deviation of error and its confidence interval of 90%. Unit: [msec]

Pre phoneme_ Current phoneme	Num.	Average of error	Absolute average error	Standard deviation	A confidence interval of 90%
i_r	127	7.72	7.97	6.73	18.79
r_o	162	**−51.54**	51.54	14.75	**−75.80**
o_SIL	284	−40.17	40.65	18.22	−70.14

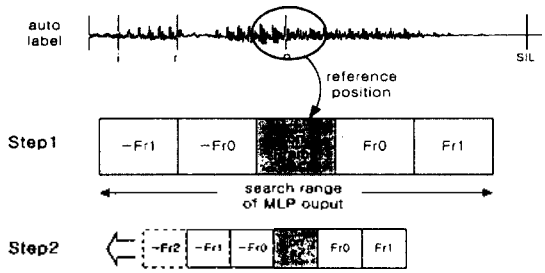

그림 2. /r/_/o/의 음운환경에 대한 경계 결정 과정

Fig. 2 The postprocessing procedure of decision boundary for /r/_/o/ phonetic environment.

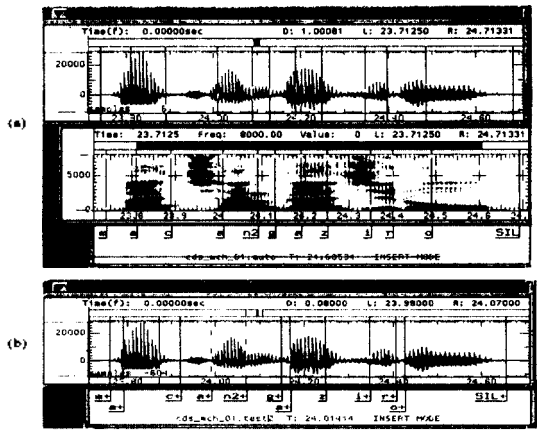

그림 3. 자동 분할기의 오류의 예와 후처리후 분절 결과: "마찬가지로" (a) '모음+자음+모음' (e.g., /i/_/r/_/o/)의 자동 분절 예, (b) (a)의 후처리후 분절 결과. '+'기호로 자동 분절 결과와 구분함

Fig. 3 An example of error by automatic segmentation system and its result after postprocessing. (a) An example of an auto label in the case of 'vowel+consonant+vowel', (e.g., /i/_/r/_/o/), (b) result of the new label from auto label, indicated by '+' sign.

V. 실험 및 결과

실험에 사용된 음성 데이터는 남성 화자가 발성한 어절 단위의 고립단어로, 음소 31,858개로 이루어진 총 3,010

개의 어절이다. 수작업에 의해 분할된 5,366개의 음소(505 word phrase)는 훈련 데이터로 사용되고 자동 분할된 음소 26,492개(2,505 word phrase)는 테스트 데이터로 사용하였다.

5.1 분절 성능에 따른 분석

MLP 후처리 적용 후 음소 분할 성능과 자동 분할기의 성능을 비교하기 위해 frame accuracy를 표 6에 나타내었다. 여기서 on frame accuracy, 1 frame accuracy 및 2 frame accuracy는 수작업된 경계로부터 오류 범위가 각각 ±5ms이내, ±15ms이내, ±25ms이내로 분할됨을 의미한다. 자동 분할기의 분할율과 후처리에 적용된 후의 분할율을 비교했을 때, 1 frame accuracy에서는 약42%에서 70%로, 2 frame accuracy에서는 약 60%에서 80%로, 약 25%의 성능향상을 보였다. 이는 제안된 후처리기로 자동 음소 분절 오류의 범위를 줄일 수 있음을 보이는 것이다.

자동 분할기가 수작업보다 일관성을 가지는 장점을 가지나 큰 오류를 포함하여 합성 단위로 직접 사용되기에 부적합하다. 따라서 이러한 큰 오류는 MLP를 이용한 후처리기로 경계 위치를 세밀하게 이동함으로써 오류의 범위를 줄일 수 있었고, 또한 위 결과는 합성단위 자동 생성에 기여할 것이다.

그러나, 제안된 방식에 여전히 포함되는 몇가지 오류가 존재한다. 첫째로, MLP탐색 구간은 자동 분절 결과를 기준으로 설정되는데, 이때 큰 오류를 포함한 자동 분절 경계가 기준인 경우 후처리로 보정할 수 있는 범위를 넘어선다. 그림 4에서 음소 /we/의 정확한 위치에 임계값 이상인 MLP 출력값이 존재하지만 자동 분절 결과의 큰 오류로 인해 음소 경계 보정에 실패한 결과를 보인다. 둘째로, 후처리의 음소 경계 결정 과정 step 2에서 탐색 구간 확장을 위해 통계 자료가 등록된 데이블로부터 음운 환경에 대한 정보를 가져오는데, 이때, 자동 분할기의 내재된 grapheme-to-phoneme 오류(i.e., transcription error : /ㅎ/ 생략 오류, /ㄴ/ 음소의 /ㅇ/ 또는 /ㅁ/으로 동화 등.)로 인해 잘못된 음운 환경 정보를 가져오면서 발생하는 오류를 포함하게 된다.

따라서 큰 오류를 포함하는 음운환경에 대한 충분한 분석이 후처리에 적용되거나, 음성학적 지식이 부분적으로 경계 결정에 이용된다면 자동 분절기의 성능 향상에 기여할 것이다. 그러나 자동 분할기의 내재된 transcription 오류는 감안해야 한다.

표 6. 프레임 분할율 Unit: [%]

Table 6. The frame accuracy of segmentation Unit: [%].

		Train data	Test data
Before Postprocessing	On frame	19.19[1,030/5,366]	18.47[4,892/26,492]
	±1 frame	41.52[2,228/5,366]	42.16[11,168/26,492]
	±2 frame	60.19[3,230/5,366]	60.09[15,920/26,492]
After Postprocessing	On frame	44.61[2,394/5,366]	**42.77**[11,330/26,492]
	±1 frame	73.37[3,937/5,366]	**70.80**[18,756/26,492]
	±2 frame	82.84[4,445/5,366]	**80.38**[21,295/26,492]

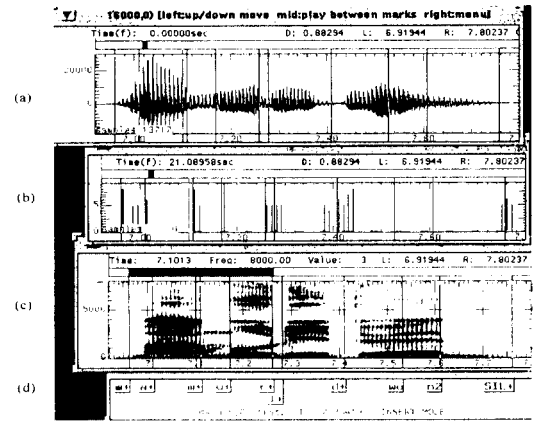

그림 4. 후처리 결과, "마무리된", (a) 음성 파형, (b) 임계값 이상의 MLP 출력값, (c) (a)의 스펙트로그램, (d) 후처리 후 분절 결과, '+'기호로 자동 분절결과와 구분함

Fig. 4 The result of postprocessing, (a) original speech waveform, (b) MLP output greater than threshold, (c) the spectrogram of (a), (d) the result of new label from auto label, indicated by '+' sign.

5.2 음운 환경에 따른 오류 향상의 분석

표 3, 4에서 본 것과 같이 음운환경 변화에 따라 자동 분절 경계 위치가 일정한 방향으로 치우쳐 분할되는 경향을 가진다. 이는 주어진 음소의 고유한 음향적 특징이 전후 음소의 영향 즉, 문맥적 영향(context influence)으로 음소 정보(phonetic information)가 중첩되고, 음소간 경계에서 특징이 상쇄되기 때문에 나타나는 현상이다. 그러나 기존 자동 분할기에 제안된 후처리 방식이 적용되었을 때, 표 3, 4에서 제시된 음운 환경에 대해 평균 오류가 표 7, 8에서처럼 3~15ms이동됨으로써 많은 향상을 가져왔다.

표 7. '[모음, 비음, 유음]+자음' 음운 환경: 후처리 전과 후처리 후의 자음의 평균 오류 비교. Unit: [msec]

Table 7. Case of '[vowel, nasal (/N/, /m2/, /n2/) or liquid (/r2/)]+consonant: The comparison of the average error of consonant before postprocessing and after postprocessing. Unit: [msec]

Postprocessing / Phoneme	Before	After
r (liquid)	17.83	8.27
b (plosive)	25.51	15.06
d (plosive)	20.35	12.08
g (plosive)	22.04	12.15
S (fricative)	18.99	5.76
Z (fricative)	12.95	4.29
D (plosive)	35.35	25.83
Z (fricative)	22.26	17.14
C (affricate)	11.87	8.91
p (plosive)	23.31	17.97
t (plosive)	23.31	16.96

표 8. '자음+모음'의 음운환경: 후처리 전과 후처리 후의 모음의 평균 오류 비교 Unit: [msec]

Table 8. Case of 'consonant+vowel': The comparison of the average error of vowel before postprocessing and after postprocessing. Unit: [msec].

Postprocessing / Phoneme	Before	After
r (liquid)	-29.81	-5.95
Z (fricative)	- 19.83	- 0.36

5.3 절대 오류 측정

성능 비교를 위해 수동 분절 위치와 자동 분절 위치와의 차이인 절대 오류(|*Hand label position-Auto label position*|)와 후처리 보정된 분절 위치와의 차이인 절대 오류(|*Hand label position-New label position*|)를 비교해 보면 21.6 msec에서 13.1 msec로 약 39% 향상되었다.

VI. 결 론

본 논문은 일정 범위의 오류 및 일관성 있는 오류를 포함하는 기존 음성 번역 시스템의 자동 분절된 결과를 MLP학습을 통해 경계 오류를 보상할 목적으로 제안되었다. 기존의 자동 음성 분할 및 레이블링 시스템은 일관성 있는 판단기준에 의해 음소 단위로 자동 분할함으로써, 대량의 합성용 운율 데이터 베이스 구축 및 음성 인식 시스템의 인식 성능을 향상에 기여할 수 있다. 그러나 자동 분절 결과에 포함되는 경계 오류로 인하여 자동 분절 시스템의 성능 향상에 대한 연구가 필수적이다. 따라서 이러한 연구의 일환으로 제안된 방법을 기존 시스템의 후처리에 도입한 결과, 자동 분할기의 성능에 비해 약 25%의 frame accuracy향상과 39%의 절대 오류 향상을 보였다.

그러나 기준이 되는 자동 분절 경계의 큰 오류 때문에 후처리로 보상할 수 있는 범위를 넘어서기도 하였다. 따라서, 이러한 국부적인 음소 분할 오류의 Phonetic knowledge를 후처리에 도입하거나, 음운 환경에 따라 다른 특징 파라메터를 적용한다면 자동 분절의 오류를 더욱 최소화할 수 있을 것이다. 이로써 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 뿐 아니라, 향후 낭독체 또는 대화체 등 문장 단위의 데이터 베이스에 제안된 후처리 방식을 적용한다면 합성용 운율 DB 및 합성 단위 자동 생성기술에 적용할 수 있을 것이다.

현재 특징 음운 환경에 강한 특징 파라메터에 대한 분석이 이뤄지고 있으며, 자유 발화 음성(Spontaneous Speech)에 대한 후처리 적용 실험이 진행되고 있다. 최종적으로, 제안된 후처리기가 도입된 분절 결과를 이용한 합성음의 평가 및 분석을 통해 그 문제점과 유용함에 대해 기술할 것이다.

참 고 문 헌

1. T. Svendsen and Frank K. Soong, "On the Automatic Segmentation of Speech Signals," *Proc. ICASSP*, pp.77-80, 1987.
2. A. W. Black & Nick Campbell, "Optimizing Selection of Units from Speech Databases for concatenative Synthesis," *EUROSPEECH*, pp.581-584, 1995.
3. 김상훈, 이항섭, 김희린, "운율 분석용 DB작성을 위한 자동 레이블러의 성능 평가 및 유용성," *SICOPS96* SESSON 3.6, 1996.
4. 박훈재, 김상훈, 정재호, "합성단위 자동생성에서의 오류에 강인한 합성단위 연결구간 결정 방법," 제10회 신호처리합동학술대회 논문집, pp.275-278, 1997.
5. 김상훈, 이정철, 강동규, 이영직, "대용량 운율 음성 데이터를 이용한 자동합성방식," 제15회 음성통신 및 신호처리 워크샵, pp.87-92, 1998.
6. Victor W. Zue, "The Use of speech knowledge in automatic speech recognition," *Proceeding of the IEEE*, pp. 1602-1615, 1985.
7. Y. Suh and Y. Lee, "Phoneme Segmentation of Continuous Speech Using The Multi-Layer Perceptrons," *Proc. of ICSLP*, pp.1293~1296, 1996.
8. John R. Deller, Jr., John G. Proakis, and John H.L. Hansen, *Discreet-Time Processing of Speech Signal*, pp. 117-137. Macmillan Publishing Company, 1993.
9. J.P. Marten and L. Depuydt, "Broad phonetic classification and segmentation of continuous speech by means of neural networks and dynamic programming," Speech communication, pp.81~90, 1991.
10. Ronald A. Cole and Lily Hou, "Segmentation and Broad Classification of Continuous Speech," *Proc. ICASSP*, pp. 453-456, 1988.
11. David B.Grayden and Michael S. Scordilis, "Phonemic Segmentation of Fluent Speech," *Proc. ICASSP*, pp.73-76, 1994.
12. 서울 대학교 생산기술 연구소, "한국어 연속 음성의 음소군 분류에 관한 연구," 한국 전자 통신 연구원 최종 연구 보고서, 1989년.

▲박 은 영(Eun-Young Park) 1973년 9월 29일생

1996년 2월 : 인하대학교 전자공학과 학사

현재 : 인하대학교 전자공학과 석사과정

※주관심분야 : 음성합성, 음성인식

e-mail : bokyung@ee.inha.ac.kr

▲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

e-mail : ksh@zenith.etri.re.kr

▲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, 신호처리 연구실, 연구원

1991년 ~ 1992년 : 미국 AT&T Bell Labolatories, 음성 신호처리 연구실, 연구원

1992년 ~ 현재 : 인하대학교 공과대학 전자공학과, 현(부교수)

e-mail : jhchung@inha.ac.kr