접합 왜곡의 최소화 과정이 포함된 음성합성기

Text-to-Speech Synthesizer with the Process of Minimizing Concatenation Distortion

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

※본 논문은 97년도 인하대학교 연구비 지원에 의하여 수행되었습니다.

요 약

대용량의 음성합성용 데이터베이스를 용이하게 구축하기 위해 음성인식 시스템을 이용한 음소 경계 분할이 이루어지고 있다. 그러나 자동 분할 결과를 직접 이용하여 합성음을 생성할 경우 음소 경계 에러로 인하여 접합 왜곡이 많이 발생하게 된다. 이러한 문제를 해결하기 위해서, 본 연구에서는 단위 접합시 경계 에러를 고려하여 적합한 접합 위치를 찾고자 하였다. 여기서 적합한 접합 위치는 스펙트럼의 불연속이 최소화된 접합점을 의미한다. 합성음에 대한 MOS(Mean Opinion Score) 테스트와 스펙트로그램(spectrogram)의 모양을 비교하므로써 제안된 방법의 성능을 평가하였다.

제안된 방법은 두 단계로 이루어져 있다. 첫째, 레퍼런스 패턴(reference pattern)과 두 개의 테스트 패턴(test pattern)을 선택하는 단계와, 둘째, 앞과 뒤 테스트 패턴 사이의 적합한 접합위치를 찾는 단계이다. 본 연구에서는 패턴 사이의 스펙트로그램 비교를 위해 켑스트럼(cepstrum) 파라미터와 패턴 분류기(pattern classifier)인 DTW(Dynamic Time Warping) 알고리즘을 사용하였다.

제안된 알고리즘을 평가한 청취 테스트의 결과에서 제안된 알고리즘을 적용하여 합성된 합성음의 음질이 자동 분절로 생성된 단위를 그대로 이용한 경우의 음질보다 우수함을 보였다.

ABSTRACT

Automatic segmentation using speech recognition system has been used for making a large size of speech synthesis data base. However, when the automatic segmentation were applied to make synthesis speech, significant concatenation distortion happens due to phoneme boundary error. To solve this problem, in our study we try to choose an appropriate concatenating position for automatically generated synthesis units which would have errors on their boundaries. Here we define the appropriate concatenating position as the place at where the spectral discontinuity between two units concatenated is minimum. We have performed MOS(Mean Opinion Score) tests and analyzed the shape of spectrograms to evaluate our proposed algorithm.

The whole procedure consists of two steps. The first step is to determine reference pattern and two test patterns. And the second step is to choose the appropriate concatenating position between the preceding and following test patterns. In our study we use cepstrum parameter and DTW(Dynamic Time Warping) pattern classifier for comparing the shape of spectrograms of the patterns.

Our test results have shown that the quality of synthesized speeches with proposed algorithm is superior to that of synthesized speeches generated by applying automatic segmentation only.

I. 서 론

음성합성 시스템의 성능은 전하고자 하는 정보를 얼마나 정확한 발음으로, 자연스럽게 합성음을 만들 수 있는 가에 달려 있다. 정확한 발음은 이해도, 선명도와 직결되며, 합성음이 명확하지 않을 경우 의사 전달이 불편하게 될 것이다. 또한 합성음이 자연스럽지 않을 경우, 인간은 그 음성에 대해 거부감을 느끼게 된다. 따라서 자연성과 명료도는 합성기가 실생활에서 사용될 수 있는지 결정하는 중요한 요소임을 알 수 있다. 합성음의 자연성은 말소리의 길이(지속 시간), 억양, 세기, 악센트, 휴지길이 등 운율 파라미터에 의해 좌우되며, 명료도는 합성단위 및

* 범일음성처리 연구소
** 한국전자통신연구원 음성언어처리 연구실
*** 인하대학교 전자공학과

접수일자 : 1998년 3월 1일

합성방식과 밀접한 관련이 있다. 특히 합성단위를 연결하여 합성하는 방식(Concatenating synthesis system)에서는 합성단위(Synthesis units)의 선택이 매우 중요하다. 이러한 합성방식에서 명료도를 저해하는 주요 요인은 합성단위간 연결시 음성신호간 스펙트럼 왜곡(Spectral mismatch)과 운율 조절을 위해 합성단위의 음성신호를 과도하게 가공하는 데서 기인한다. 따라서 명료한 합성음을 생성하기 위해서는 합성단위간 연결시 음성신호간 왜곡을 줄이는 방식인 모음의 안정 구간에서 연결하는 다이폰 또는 반음절 단위를 사용하거나, 연결점의 개수를 줄이기 위해 음소보다 큰 음절이나, 최장일치 방식으로 합성단위를 선정하는 방법을 사용한다.[1] 또한 합성단위의 음운 환경을 제약하지 않고 가능한 다양한 음운환경이 고려되도록 합성단위(예:triphones)를 작성하기도 한다. 그리고 합성단위 선정 방식에서는 원 음성신호의 신호처리(가공)를 최소화하기 위해 대량의 데이터로부터 복수 후보의 합성단위 중 가장 적절한(appropriate) 단위를 선정하기도 한다.[2][3]

이와 같이 보다 큰 합성단위 선정, 복수 후보 합성단위 구축 및 다양한 음운 환경을 고려하기 위해서는 대량의 합성 데이터베이스 구축이 필요하고 이를 위해 합성단위 제작의 자동화 과정이 매우 중요한 기술이 된다. 그러나 아직 음성 인식기를 이용한 자동 분절의 결과가 합성단위로 바로 사용될 수 없고, 수동으로 자동 분절의 오류를 수정해야 하는 과정이 필요하다. 이러한 수정 작업은 여전히 시간이 많이 소모되는 일이다.

본 논문에서는 대용량 합성 데이터베이스를 효율적으로 구축하기 위해 전처리로 음성인식기를 사용하여 음소분할을 자동으로 수행하고, 음성인식기의 음소분할 오류에 대해 후처리에서 보정할 수 있도록, 자동 분절의 오류를 감안한 합성단위 연결 구간 결정 방식을 제안한다. 근래 음성인식 시스템의 성능 향상으로 음성합성 분야에서도 인식기를 이용하여 합성단위를 자동 추출하는 연구가 활발히 진행되고 있으며 본 연구도 이러한 노력의 일환으로써 대용량 합성 데이터베이스 구축의 자동화와 이로부터 합성음질의 향상을 이루는 것이 연구의 목적이다.

본 논문의 구성은 다음과 같다. 서론에 이어 제 2장에서는 기존 음성합성 시스템을 완성하기 위한 작업 과정을 기술하고, 3장에서는 본 연구에서 사용하는 켑스트럼 파라미터와 DTW 알고리즘의 기본 개념에 대하여 설명하고 이를 이용한 합성단위 연결 구간 결정 방식을 구체적으로 설명한다. 4장에서 제안된 알고리즘을 이용한 실험과 결과를 기술하고 5장에서는 본 논문의 결론과 앞으로의 연구 방향을 제시하면서 논문을 마무리 짓는다.

II. 음성합성 시스템

2장에서는 일반적인 음성합성 시스템(TTS:Text-to-Speech System)에 대하여 설명한다.

2.1 합성 데이터베이스 구축

음성 데이터는 우선 자동 레이블링 시스템을 이용하여 모든 음성 데이터를 음소 경계로 분할하고, 수작업으로 음소 경계를 보정하여 데이터베이스를 구축하였다.[4][5]

본 연구에서 사용한 자동 레이블링 시스템은 한국 전자통신 연구원에서 개발한 음성언어 번역시스템으로 대화체 음성인식 및 한/일, 한/영 번역을 목적으로 하고 있다. 현재 여행 영역에서 약 5,000 단어를 인식할 수 있으며, HMM (Hidden Markov Model) 인식 알고리즘을 사용하고, 단어 인식률이 약 70% (perplexity=110[1])에 이르고 있다. 그림 2.1은 자동 레이블러의 분절 오류의 예를 보여준다. 레이블링 시스템의 성능 비교를 위하여 수동으로 레이블링한 결과와 비교하였으며 비교는 수동에 의한 레이블링 위치와 자동에 의한 레이블링 위치를 비교하여 그 차이의 절대치를 오류로 하였다. 자동 레이블링 오류가 30msec 이내인 결과를 고려할 때, 고립단어인 경우 남자의 경우 63%, 낭독체 문장인 경우 82%, 대화체의 대화자인 경우 78%, FM radio news는 87%의 성능을 보여 주고 있다[6].

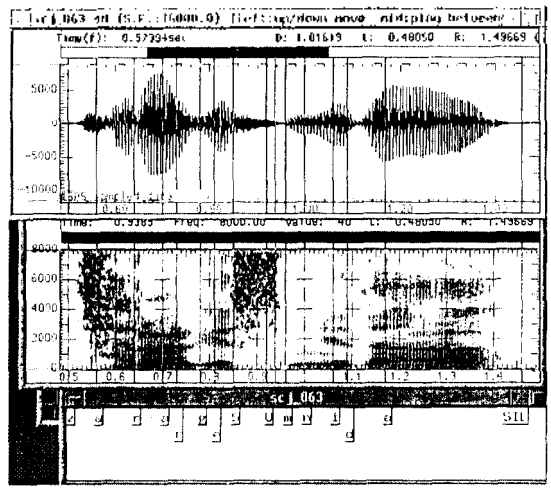


그림 2.1 자동 레이블링 오류 예("잘알겠습니다")
Fig. 2.1 Example of an error by automatic.

2.2 합성단위 선정[7]

본 연구에서는 합성단위로 트라이폰을 사용한다. 트라이폰 합성단위는 음성인식에서 사용하는 단위와 동일하다. 즉, 음소를 기준으로 좌우 음운 환경이 다르면 하나의 트라이폰 합성단위는 다양한 음운 환경을 가지고 있어서 기존 음운 환경이 제약된 단위(예 : CDU:Context Dependent Units)보다 합성음의 명료도 및 자연성을 향상시킬 수 있다. 그러나 한국어에서 발생하는 트라이폰 개수는 약 6만여개 정도 발생하며 트라이폰 합성단위를 구축하기 위해서는 자동화 과정이 매우 중요한 기술이 된다. 음소 경계 레이블링 작업이 완료된 데이터베이스 내에서 발생하는 모든 트라이폰에 대한 정보를 사전으로 등록한 후, 합성음을 생성하는 과정에서 필요한 트라이폰

1) 하나의 단어를 인식하는데 찾는(search) 평균 단어 개수

을 찾을 때 이 사전을 이용하여 음성 데이터베이스의 트라이폰을 읽게 된다.

그림 2.2에 2장에서 기술한 합성 시스템의 제작 과정을 간략히 나타내었다.

그림 2.2 기존 합성기 구현 과정
Fig. 2.2 The process of conventional synthesizer implementation.

III. 경계 오류에 강인한 합성단위 연결구간 결정 방법

3장에서는 먼저 수동 보정 작업의 문제점을 기술한 후, 자동 분절의 오류를 감안한 합성단위 연결구간 결정 알고리즘을 설명한다.

3.1 수동 보정 작업의 문제점과 경계 에러로 인한 접합 왜곡[3][8][9]

일반적으로 수동 음소 분절 작업은 다음과 같은 문제점을 지닌다. 첫째로 이 과정은 스펙트로그램 판독 및 반복되는 듣기 평가를 통해 이루어지므로 매우 지루한 작업일 뿐만 아니라 많은 시간이 소요되게 된다. 둘째로 수작업에 의한 음소 분할은 높은 수준의 음성학적 지식을 요하므로, 소수의 음성학 전문가에 의존할 수밖에 없다. 셋째로 음소 경계 선정을 위한 구체적인 판단기준을 미리 정해놓더라도 상당 부분의 경우 주관적인 판단을 피할 수 없으며, 이에 따라 음소 경계 분절 과정에서의 일관성이 보장되지 못한다. 음소 분절 작업이 자동으로 수행될 수 있다면 위에서 언급한 문제들이 해결될 수 있다. 그러나 자동 레이블러의 성능이 만족할 만한 수준이 아니므로(2장 1.1절 참조) 이 데이터를 합성기에 그대로 사용한다면 경계에러로 인한 접합 왜곡이 발생할 것이다.

본 논문에서는 접합 왜곡의 세 가지를 정의하였다. 그림 3.1의 (a)는 생략 왜곡, (b)는 삽입 왜곡을, 그리고 (c)는 스펙트럼 불연속 왜곡을 나타낸다. (a), (b)의 실선 화살표는 정확한 음소 경계를 이용하여 접합하는 경우로, 생략, 삽입 왜곡이 발생하지 않는다. 그러나 점선 화살표는 에러가 있는 음소 경계를 이용하여 접합하는 경우를 나타내는데, (a)는 앞쪽 음소와 뒤쪽 음소의 경계 부분이 생략되는 왜곡이 발생하고, (b)는 두 음소의 경계 부분이 추가되는 삽입 왜곡이 발생한다. 또한 (c)는 부적절한 접합점의 위치로 인하여 발생하는 스펙트럼 불연속의 왜곡을 나타낸 것으로 앞쪽 단위와 뒤쪽 단위 사이의 포먼트의 불연속이 발생한 경우이다. 부적절한 접합점으로 인하여 발생하는 이러한 왜곡으로 합성 음질의 저하가 초래되므로, 이것은 반드시 해결되어야 할 문제이다.

3.2 켑스트럼 계수와 DTW(Dynamic Time Warping)[10][11][12][13][14]

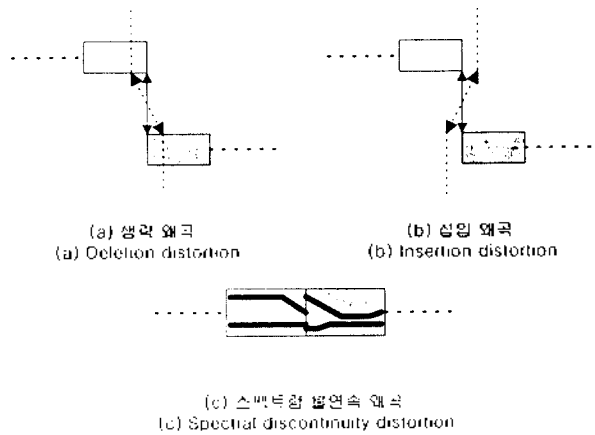


그림 3.1 접합 왜곡
Fig. 3.1 Concatenation distortion.

켑스트럼 계수는 호모몰픽 필터링에 의한 역컨볼루션(homomorphic deconvolution)으로 생성되며, 음성 신호의 경우 켑스트럼 영역에서 음원 신호에 대한 정보와 성도의 임펄스 응답에 관한 정보가 서로 겹치지 않는 성질이 있어서 음성 신호에 관한 좋은 분석법으로 알려져 있다. 켑스트럼 영역에서 성도 부분의 정보는 아래쪽 영역에, 그리고 음원 신호의 정보는 위쪽 영역에 각각 분포하게 되므로 주로 성도 부분의 정보를 이용하게 되는 음성의 패턴 인식에서는 12차~16차 정도의 켑스트럼 계수를 취하여 사용하는게 보통이다. 켑스트럼 계수는 LPC(Linear Predictive Coding) 계수에 비해서 통계적 특성이 좋고, 또한 간단한 유클리드 거리(Euclidean distance)만으로도 두 신호의 스펙트럼 왜곡(Spectral distortion)을 추정할 수 있는 등의 장점 때문에 음성 인식에서 주로 사용하는 파라미터가 되었으며, 실제로 인식률 면에서도 LPC 계수를 사용하는 것보다 나은 성능을 보이고 있다. LPC-켑스트럼 계수는 켑스트럼의 성질을 충분히 나타내면서 적은 계산량으로 구할 수 있으므로 본 연구에서는 LPC-켑스트럼 계수를 음성 특성 파라미터로 사용하였다.

음성 인식의 경우에 보통의 패턴 인식과는 다른 특징이 존재하는데, 이것은 바로 음성은 근본적으로 시간에 따라 변하는 신호라는 점과 시간축상에서 부분적으로 확장 또는 수축(즉, time-warped)되어진다는 특성이 있다. DTW 알고리즘은 두 음성을 시간적으로 정렬하면서 거리를 계산하는 음성 인식 알고리즘이다. DTW는 본 연구에서 사용하는 음성 데이터에 적합하고 합성단위간 접합점을 찾는데 이용하는 레퍼런스 패턴과 테스트 패턴간의 거리를 계산하는데 매우 적합한 알고리즘이라 할 수 있다.

3.3 제안된 접합 방법[3][9][15]

그림 3.2는 제안된 알고리즘의 개요도로서 크게 나누어 두 과정으로 이루어져 있다. 접합하여 만들고자 하는 단위와 같은 스펙트로그램 형태를 지닌 레퍼런스 패턴(reference pattern), 접합하려는 두개의 테스트 패턴(test

pattern)을 선택하는 과정과, 선택된 두 개의 테스트 패턴을 여러 위치에서 접합하면서 후보 접합 테스트 패턴을 만들고 레퍼런스 패턴과 스펙트로그램이 가장 유사한 후보 테스트 패턴을 선택하는 과정이다. 레퍼런스 패턴과 테스트 패턴은 자동 레이블링을 거쳐 생성된 데이터베이스로부터 선택되어진다.

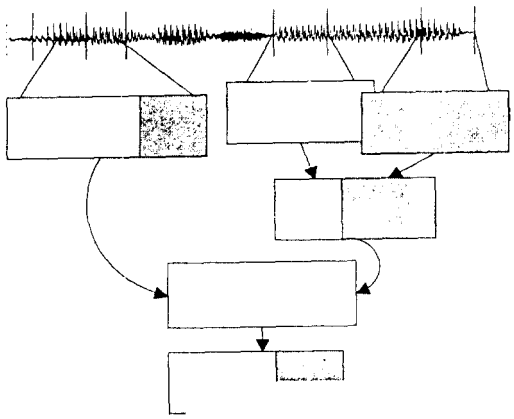


그림 3.2 제안된 방법의 개요도
Fig. 3.2 Overview of proposed method.

첫 번째 과정은 그림 3.3에 나타내었다. 접합하여 만들고자 하는 단위와 스펙트로그램 형태가 같은 레퍼런스 패턴을 선택한 후 이것을 참조하여 접합하려는 두 개의 테스트 패턴을 선택한다. 레퍼런스 패턴의 스펙트로그램과 가장 유사한 테스트 패턴을 선택하기 위하여 테스트 패턴 추출 창을 자동 레이블링된 음소 경계를 중심으로 앞, 뒤로 이동해 가며 테스트 패턴 후보를 추출하고, 이것들을 레퍼런스 패턴과 비교하면서 스펙트로그램의 형태가 가장 유사한 것을 선택한다. 일단 앞 테스트 패턴이 선택되어지면, 이에 상응하는 뒤 테스트 패턴도 같은 방법으로 선택되어진다.

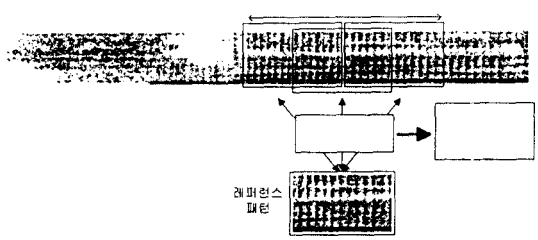


그림 3.3 테스트 패턴 선택 방법
Fig. 3.3 A method to choose test pattern.

그림 3.4는 두 번째 과정을 나타내는데 선택되어진 앞, 뒤 테스트 패턴 사이에 적합한 접합점을 찾는 것이다. 먼저 두 테스트 패턴의 접합 가능한 구역을 정하고, 이 구역을 여러 개의 프레임으로 나눈다. 앞, 뒤 테스트 패턴을 각각 I와 J 프레임으로 나눈다고 가정하면 접합된 패턴 SP_{ij}, $\{1\le i\le I,\ 1<j\le J\}$ 중에서 레퍼런스 패턴과 비교되어 가장 유사한 것이 선택되어진다. 선택되어진 후보 접합 테스트 패턴이 앞쪽 테스트 패턴($1\le i\le i^*$ 프레임)과 뒤쪽 테스트 패턴($j^*\le j\le J$ 프레임)으로 이루어지는 접합 테스트 패턴이라면 접합 위치는 i^*, j^* 번째 프레임이 되는 것이다.

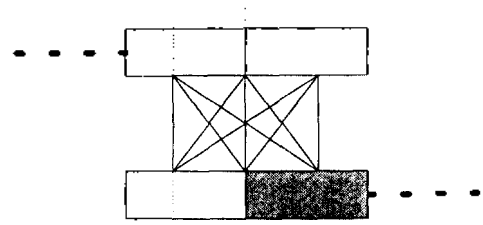


그림 3.4 두 테스트 패턴을 접합하는 과정
Fig. 3.4 The process of concatenating the two test patterns.

제안된 알고리즘을 이용하여 합성음을 생성하는데 사용하는 음성 분석, 특징 파라미터 추출 및 레퍼런스와 테스트 패턴의 선택은 다음과 같다.

Sampling rate	16kHz
A/D 양자화 해상도	16bits
분석 구간	300samples
분석 주기	10samples
분석 창	Hamming window
특징 파라미터	20차 LPC-cepstrum
레퍼런스, 테스트 패턴 크기	600samples
테스트 패턴 추출 이동 주기	10samples
레퍼런스, 테스트 패턴 추출에 사용된 창	Rectangular window

그림 3.5는 제안된 알고리즘을 2장에서 기술한 합성기에 적용하여 합성기의 제작 과정을 개선한 것을 나타낸다. 수동 보정 작업이 제외되고 제안된 적합 방법이 첨가되었다.

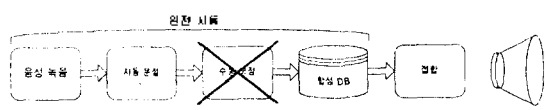


그림 3.5 개선된 합성기 제작 과정
Fig. 3.5 The process of implementation of improved synthesizer.

IV. 실험 및 결과

기존의 방법과 제안된 방법으로 합성음을 생성하고 MOS 테스트 방법을 사용하여 제안된 알고리즘의 성능을 평가하였다. 각 방법으로 생성된 합성음을 들려주면서 5점에서 1점의 점수로 합성음을 평가하는 방법으로 수행하였

다. 청취 평가는 접합점에서의 스펙트럼 왜곡에 중점을 두어 이루어졌다.

4.1 청취 테스트

MOS 테스트를 수행하기 위하여 임의로 7개의 합성 문장을 만들고 각 합성 문장에 대하여 다음의 세가지 방법으로 합성음을 생성하였다.

(A) 음소 경계 보정 작업을 거친 데이터를 이용할 경우
(B) 자동 분절된 데이터를 그대로 이용할 경우
(C) 자동 분절된 데이터에 본 연구에서 제안된 방법을 적용한 경우

성능 평가에 참가한 인원은 음성 분야에 비전문적인 청취자 10명이다. 표 4.1은 각 문장에 대한 방법 (A), (B), (C)의 MOS 테스트 결과를 나타낸다. 각 문장의 각 방법에 대한 평균과 각 방법에 대한 문장 전체의 평균, 표준 편차를 나타내었다.

표 4.1 합성음 청취 테스트 결과
Table 4.1 The result of the test of synthesized speech quality.

문장 / 방법	1	2	3	4	5	6	7	평균	표준편차
(A)	3.6	3.1	4.2	4.1	3.6	4.4	3.8	**3.84**	0.93
(B)	3.4	3.9	3.5	3.1	3.2	3.1	2.9	**3.30**	0.95
(C)	3.2	3.4	3.5	3.1	3.7	3.5	3.5	**3.44**	0.88

수동 보정 작업을 거친 데이터를 이용한 경우 (A)가 평균 3.84로 가장 높은 점수를 얻었다. 그리고 자동 분절된 데이터에 제안한 알고리즘을 적용하여 접합한 경우 (C)가 3.44로써, 자동 분절된 데이터를 그대로 이용하여 접합한 경우 (B)의 3.30보다 높은 점수를 얻었다. 이 결과는 자동 분절에 의한 경계 에러를 제안된 방법을 이용하여 보상하였음을 나타낸다. 즉, 접합점에서의 왜곡을 줄였고 따라서 합성음 전체적으로 음질이 향상되었음을 나타낸다.

4.2 스펙트로그램 비교를 통한 결과 분석

그림 4.1과 4.2는 (A), (B), (C) 방법으로 생성된 합성음의 일부를 자세히 나타낸 것으로써, 그림 4.1은 제안된 방법으로 적합한 접합점을 찾은 예이고, 그림 4.2는 적합한 접합점을 찾지못한 예를 나타낸다. 그림 4.1(a)는 음소 /a/와 /n/가 방법 (A), (B), (C)로 접합된 결과를 나타낸 것이다. 첫 번째와 세 번째 그림을 보면 접합점에서 스펙트럼의 불연속이 발생하지 않았고, 두 번째는 접합점에서 스펙트럼의 불일치가 발생함을 볼 수 있다. 그림 4.1(b)는 방법 (B)로 /a/와 /n/을 접합하는 과정을 나타낸다. 음성 데이터베이스 내에 존재하는 음소 경계의 에러로 인하여 접합점에서 스펙트럼의 불일치가 발생함을 볼 수 있다.

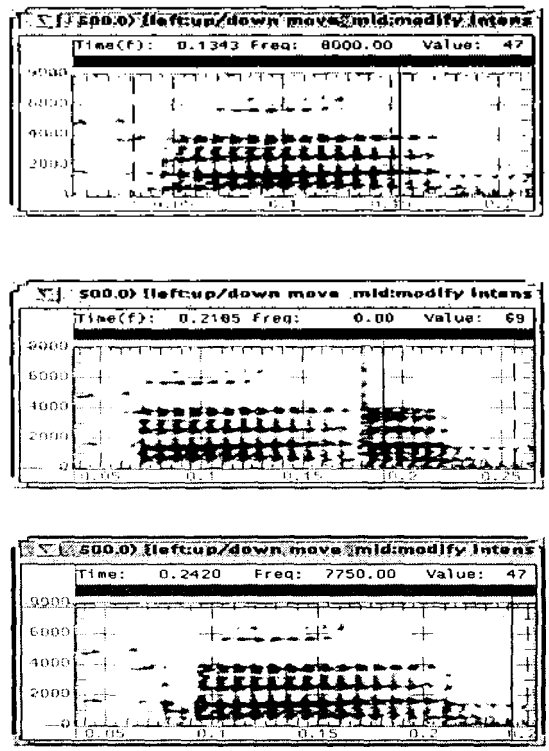


그림 4.1(a) /a/와 /n/의 접합 부분
(첫번째:방법(A), 두 번째:방법(B) 세번째:방법(C))
Fig. 4.1(a) concatenation between /a/ and /n/
(First:method(A), Second:method(B), Third:method (C)).

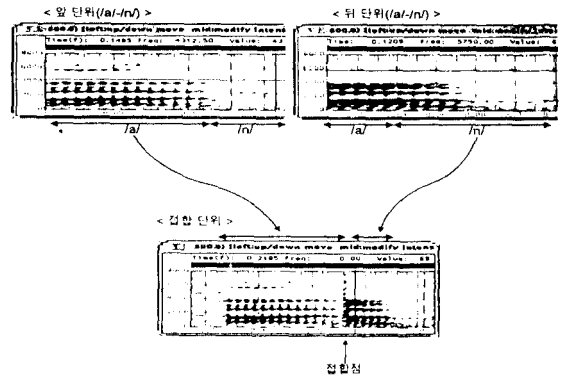


그림 4.1(b) 방법(B)로 /a/와 /n/의 접합하는 과정
Fig. 4.1(b) the process of concatenation between /a/ and /n/ using method (B).

그림 4.1(c)는 방법 (C)로 접합하는 과정을 나타내는데 미리 선택되어진 레퍼런스 패턴을 이용하여 두 음소 사이의 접합점에서 스펙트럼의 불일치가 없는 위치를 찾았다. 그림 4.2(a)는 음소/a/와 /r/가 방법 (A), (B), (C)로 접합된 결과를 나타낸 것으로써, 제안된 방법 (C)를 이용하여 접합된 결과인 세 번째 그림에서 스펙트럼의 불연속이 많이 발생하였다. 이것의 원인은 다음 그림을 보면 알 수 있다. 그림 4.2(b)는 방법 (B)로 /a/와 /r/을 접합하는 과정인데 두 접합단위의 /a/와 /r/을 경계가 모두 앞쪽 음소

/a/쪽으로 치우쳐있어서 두 단위를 붙였을 때 접합점에서 스펙트럼 불일치가 발생하지 않았다. 그러나 그림 4.2(c)를 보면 접합점의 위치를 찾는데 사용된 레퍼런스의 스펙트로그램 패턴이 두 개의 테스트 패턴과는 상이하므로 적합한 접합점을 찾지못하는 결과를 발생하였다. 생성된 합성음을 전체적으로 보면 이러한 잘못된 레퍼런스 단위의 사용으로 인한 에러가 대부분을 차지하였다. 그러므로 더욱 정교한 레퍼런스 단위의 선택 기준과 방법에 대한 연구가 제안된 알고리즘의 성능을 높이는데 가장 중요한 역할을 할 것이다.

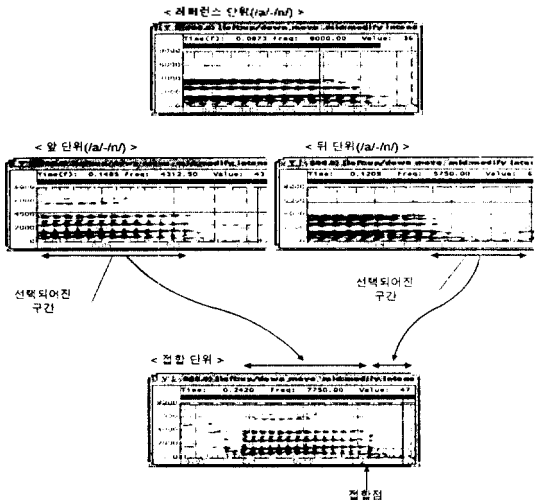


그림 4.1(c) 방법(C)로 /a/와 /n/의 접합하는 과정
Fig. 4.1(c) the process of concatenation between /a/ and /n/ using method (C).

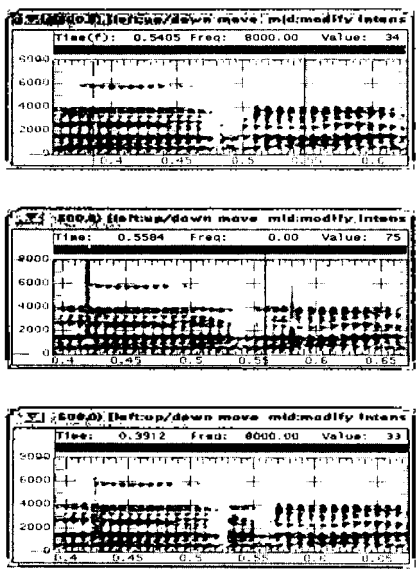


그림 4.2(a) /a/와 /r/의 접합 부분
(첫번째:방법(A), 두번째:방법(B), 세번째:방법(C))
Fig. 4.2(a) concatenation between /a/ and /r/
(First:method(A), Second:method(B), Third:method (C)).

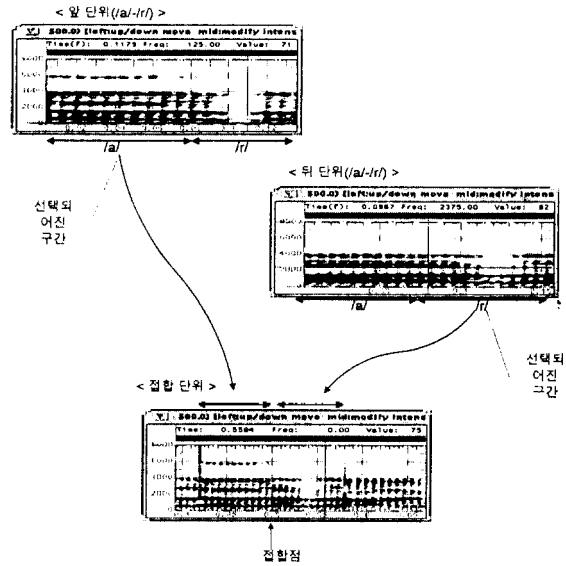


그림 4.2(b) 방법(B)로 /a/와 /r/의 접합
Fig. 4.2(b) the process of concatenation between /a/ and /r/ using method (B).

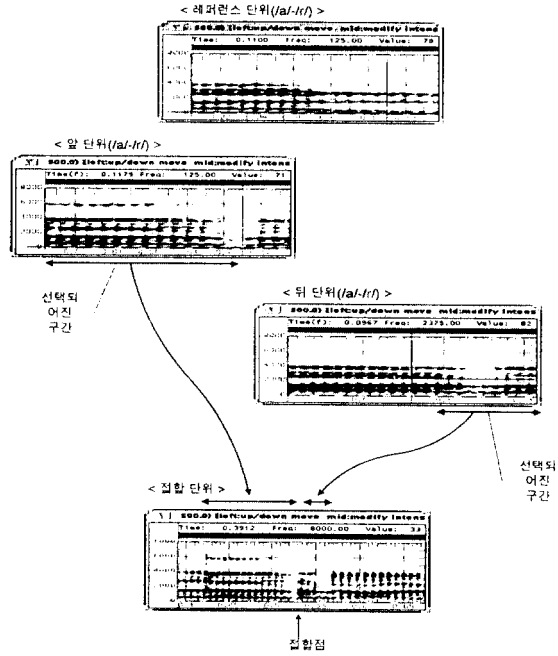


그림 4.2(c) 방법(C)로 /a/와 /r/의 접합
Fig. 4.2(c) /a/ and /r/ concatenation using method (C).

V. 결 론

본 논문에서는 자동 분절 결과를 이용할 경우 발생하는 경계 에러를 합성단위 접합 과정에서 고려하여 접합 왜곡을 최소화하는 방법을 연구하였다.

합성 음질의 향상을 위하여 보다 큰 합성단위 선정, 복수 후보 합성단위 구축 및 다양한 음운 환경을 고려하기 위해서는 대량의 합성 데이터베이스 구축이 필요하고

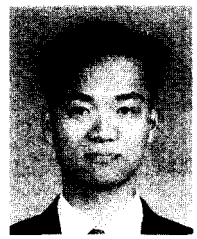
이를 위해 음성인식 기술을 이용한 자동 분절은 중요한 기술이 된다. 그러나 경계 에러로 인하여 아직 자동 분절의 결과가 합성단위로 바로 사용될 수 없고, 수동으로 자동 분절의 오류를 수정해야 하는 과정이 필요하다. 본 연구에서는 수동으로 수정하는 과정 대신에, 합성단위 접합 과정에서 레퍼런스 단위를 이용하여, 접합 왜곡이 최소가 되는 접합점을 찾는 과정을 첨가하므로 합성 데이터베이스 구축의 완전 자동화와 합성음질의 향상을 추구하였다. MOS 테스트로 제안된 방법의 성능을 평가한 결과 자동 분절된 데이터를 그대로 이용할 경우 3.30, 자동 분절된 데이터에 본 연구에서 제안된 알고리즘을 적용한 경우는 3.44의 점수를 얻어서 제안된 방법으로 접합 왜곡을 줄일 수 있음을 확인하였다.

그러나, 실험 결과에서 확인하였듯이 제안된 알고리즘의 해결되어야 할 점은 적절치 못한 레퍼런스 패턴의 사용 가능성이다. 즉, 접합하여 만들고자하는 단위의 스펙트로그램의 형태와 큰 차이가 생기는 레퍼런스 패턴을 사용하므로 적합한 접합점을 찾지 못하는 결과가 발생하였다. 이 문제를 해결하기 위하여 레퍼런스 패턴의 선택에 있어서 더욱 상세한 기준과 방법이 연구되어져야 한다. 그리고 대량의 데이터베이스에는 다양한 환경의 복수 후보단위가 발생하는데 이러한 복수개의 단위 중 접합왜곡을 최소화할 수 있는 단위 선택에 관한 연구가 이루어져야 한다.

참 고 문 헌

1. Yoshinori Sagisaka, "Speech Synthesis by Rule Using an Optimal Selection of Non-Uniform Synthesis Units," *Proc. ICASSP*, pp.697-682, 1988.
2. Alexander G. Hauptmann, "SPEAKEZ:A First Experiment In Concatenation Synthesis From A Large Corpus," *EUROSPEECH*, volume 3, pp.1701-1704, 1993.
3. A. W. Black & Nick Campbell, "Optimising Selection of Units from Speech Databases for concatenative Synthesis," *EUROSPEECH*, pp.581-584, 1995.
4. "한국어의 운율 분석 및 음운의 분절 표기에 관한 연구," 한국 전자 통신 연구소 최종연구 보고서, 서울대학교 인문대학, pp.6-31, 1993.
5. 김종신 외, "한국어 음성 DB 구축을 위한 한국어 레이블링 기준에 관한 연구," 음성 통신 및 신호처리 워크샵 논문집 제 13회, pp.250-255, 1966.
6. 김상훈, 이항섭, 김회린, "운율 분석용 DB 작성을 위한 자동 레이블러의 성능 평가 및 유용성," *SICOPS96* SESSON 3.6, 1996.
7. 고려대학교 정보통신기술공동연구소, "양질의 음성합성을 위한 최적의 합성단위 추출에 관한 연구," 한국전자통신 연구소 보고서 1993.
8. H. C. Leung and V. Zue, "A procedure for automatic alignment of phonetic transcriptions with continuous speech," *Proc. ICASSP*, pp.429-432, Apr. 1984.
9. Naoto Iwahashi, et tal, "Concatenative Speech Synthesis by Minimum Distortion Criteria," *Proc. ICASSP*, Vol.II, 1992
10. Hisashi Wakita, "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms," *IEEE Trans. On Audio and Elecroacoustics*, Vol.AD-21, No.5, pp.417-427, Oct. 1973.
11. B. S. Atal, S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, vol.50, pp.637-655, 1971.
12. L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood cliffs, N.J., 1978.
13. L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood cliffs, N.J., 1993.
14. M. Ionita, C. Burileanu, M. Ionita, "A Version of DTW Algorithm with Associated Mattix," *Proc. ICSP*, pp.433-437, 1997.
15. 박훈재, 김상훈, 정재호, "합성단위 자동생성에서의 오류에 강인한 합성단위 연결구간 결정 방법," 제 10회 신호처리 합동 학술대회 논문집, pp.275-278, 1997.

▲박 훈 재(Hun Jae Park) 1973년 6월 26일생



1996년 2월:인하대학교 전자공학과 학사
1998년 2월:인하대학교 전자공학과 석사
현재:범일음성처리 연구소 연구원
※주관심분야:음성합성, 음성인식
e-mail:jim@bic.co.kr

▲김 상 훈(Sanghun 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@dragon.inha.ac.kr