한국음향학회지 제22권 제2호 pp. 135~144 (2003)

문자-음성 합성기의 데이터 베이스를 위한 문맥 적응 음소 분할

Context-adaptive Phoneme Segmentation for a TTS Database

이 기 승*, 김 정 수**
(Ki-Seung Lee*, Jeong-Su Kim**)

*건국대학교 정보통신대학 전자공학과, **삼성전자(주) 종합기술원 휴먼-컴퓨터 인터엑티브 연구실

(접수일자: 2002년 11월 18일; 채택일자: 2003년 1월 10일)

본 논문에서는 문자-음성 합성기에서 사용되는 대용량 데이터 베이스의 구성을 목적으로 하는 음성 신호의 자동 분할 기법을 기술하였다. 주된 내용은 은닉 마코프 모델에 기반을 둔 음소 분할과 여기서 얻어진 결과를 초기 음소 경계로 사용하여 이를 자동으로 수정하는 방법으로 구성되어 있다. 다층 퍼셉트론이 음성 경계의 검출기로 사용되었으며, 음소 분할의 성능을 증가시키기 위해, 음소의 천이 패턴에 따라 다층 퍼셉트론을 개별적으로 학습시키는 방법이 제안되었다. 음소 천이 패턴은 수작업에 의해 생성된 레이블 정보를 기준 음소 경계로 사용하여, 기준 음소 경계와 추정된 음소 경계간의 전체 오차를 최소화하는 관점에서 분할되도록 하였다. 단일 화자를 대상으로 하는 실험에서 제안된 기법을 통해 생성된 음소 경계는 기준 경계와 비교하여 95%의 음소가 20 msec 이내의 경계 오차를 갖는 것으로 나타났으며, 평균 자승 제곱근 오차면에서 수정 작업을 통해 25% 향상된 결과를 나타내었다.

핵심용어: 문자-음성 합성기, 과형 연결, 스무딩, 문맥 적응 필터, 분류 회귀 나무
투고분야: 음성처리 분야 (2.4)

A method for the automatic segmentation of speech signals is described. The method is dedicated to the construction of a large database for a Text-To-Speech (TTS) synthesis system. The main issue of the work involves the refinement of an initial estimation of phone boundaries which are provided by an alignment, based on a Hidden Markov Model (HMM). Multi-layer perceptron (MLP) was used as a phone boundary detector. To increase the performance of segmentation, a technique which individually trains an MLP according to phonetic transition is proposed. The optimum partitioning of the entire phonetic transition space is constructed from the standpoint of minimizing the overall deviation from hand labelling positions. With single speaker stimuli, the experimental results showed that more than 95% of all phone boundaries have a boundary deviation from the reference position smaller than 20 ms, and the refinement of the boundaries reduces the root mean square error by about 25%.

Keywords: *Text-To-Speech synthesis, Automatic phoneme labelling, Multi-layer perceptron, Neural network training algorithm*
ASK subject classification: *Speech signal processing (2.4)*

I. 서 론

음성 합성기[1-3]란 사용자가 임의로 입력한 문장을 컴퓨터 등을 이용하여 자동적으로 음성을 생성하여 청취자에게 들려주는 시스템을 말한다. 음성 합성기는 자동안내 시스템과 같은 응용 분야에 널리 사용되고 있으며, 사람과 기계간의 의사 소통을 가능하게 하는 기술에 있어서 중요한 역할을 담당한다. 이러한 음성 합성기는 초기에 음성의 해부학적인 발생 모델을 바탕으로 하는 포만트 합성 (formant synthesis) 기법이 소개되었으며 [1], 1990년대에 접어들면서 대용량 데이터 베이스를 기반으로 하는 코퍼스 기반 문자-음성 합성기 (corpus-based Text To Speech; TTS)[2]가 제안되어 보다 인간의 음성

책임저자: 이기승 (kseung@kkucc.konkuk.ac.kr)
143-701 서울특별시 광진구 화양동 1번지
건국대학교 정보통신대학 전자공학과 1417호
(전화: 02-450-3489; 팩스: 02-3437-5235)

에 가까운 합성음을 얻게 되었다. 또한 운율 예측 (prosody prediction) 에 있어서도 통계적인 모델을 이용한 데이터 구동 (data-driven) 기법이 적용되어 보다 생동감 넘치는 합성음을 얻게 되었다.

고음질, 대용량의 데이터 베이스를 특징으로 하는 현재의 코퍼스 기반 문자-음성 합성기가 기존의 기계적인 합성음을 극복한 주된 이유로는 다양한 합성음의 합성을 가능하게 하는 방대한 데이터 베이스에 있다고 볼 수 있다[4]. 이러한 방대한 데이터 베이스의 사용이 가능하게 된 주 된 이유는 하드디스크 드라이브와 같은 저장 매체의 가격 하락, 복잡하고 방대한 음소 정보를 빠른 시간에 처리할 수 있는 고속 중앙연산처리 장치 (CPU)의 개발 등이 힘입은 바가 크다고 볼 수 있다[5].

이러한 방대한 양의 데이터 베이스를 구성하기 위한 방법으로, 직접 파형을 청취하면서 음소 경계를 구분하는 수작업 레이블링 (hand-labelling)이 주로 사용되어 왔는데, 수 100 Mbytes의 방대한 음성 데이터를 음소로 분할하는데는 아주 많은 시간과 노력이 필요하다[5]. 이러한 데이터 베이스 구성시의 어려움은 음성 합성기의 개발 속도를 지연시키며 인적 자원의 확보, 인건비 상승 등의 비용 상승 요인으로 작용하게 된다. 수작업 레이블링의 또 다른 문제점은 음소 경계의 예측이 작업자의 경험, 주관적인 판단 기준 및 작업자의 심리 상태에 영향을 받을 수 있다는 점이다. 따라서 다수의 작업자에게 데이터 베이스 구성 작업을 할당하는 경우, 음소 경계에 대한 공통성 (consistency)을 보장하지 못하게 된다.

이러한 문제점을 해결하기 위한 노력으로 Larry 등은 음성 인식에 널리 쓰이는 은닉 마코프 모델 (Hidden Markov Model; HMM)을 음소의 분할에 사용하는 방법을 제안하였다[7]. 이 방법은 음성 합성용 데이터 베이스의 작성을 목표로 제안된 방법은 아니지만, 수작업에 의한 음소 경계 정보가 전혀 주어지지 않은 상황에서, 음소 경계를 자동으로 분할할 수 있는 최초의 기법이라 할 수 있다. HMM을 기반으로 하는 음소 분할 기법은 아직까지도 음소 분할의 중요한 기법으로 남아있으며, 여기에 부가적인 처리를 통해 음소 경계의 정밀도를 높이는 방법들이 다수 제안되고 있다[6,8-10].

이로서 선형 예측 계수의 시간 상관 함수을 구하여, 상관도가 높은 영역을 음소의 안정 구간으로 간주하는 자동 레이 블링 기법이 Jan에 의해 제안되었다[6]. 이 방법은 HMM에 의해 추정된 음성 경계를 수정하는 일종의 후처리 기법 (postprocessing)이 적용된 것으로, 자동으로 얻어진 음소 경계가 수작업으로 얻어진 음소 경계와 비교하여 90%의 음소가 20 msec 이내의 경계오차를 가지며, 청취 테스트 상으로도 수작업 레이블링의 결과보다 우수한 것으로 보고되었다. 그러나 이 방법은 다이폰-기반 합성기 (diphone-based synthesis system)를 대상으로 제안되었으며, 데이터 베이스로 사용되는 음성이 음소 분할이 다소 용이한 무의미 2음절어 (non-sense 2 syllable)로 구성되어 현재의 코퍼스 기반 TTS에서 사용되는 문장 (sentence)단위 음성 데이터에 대해서는 검증되지 않은 기법이라 볼 수 있다.

Bonafonte 등에 의해 제안된 가우시안 모델을 기반으로 한 음소 경계의 수정 기법[8]은 HMM기반의 음소 분할이 전역 통계적 특성 (global probabilistic model)만을 반영한 음소 모델을 사용한다는 문제점을 고려하여, 현재 분할하고자 하는 음성의 지역 통계적 특성 (local probabilistic model)도 함께 고려한 기법이다. 이 방법을 통해 HMM기반의 음소 경계가 수작업 레이블링의 결과와 가까워짐이 실험적으로 증명되었으나, 음소 조합에 따라 성능의 편차가 비교적 심하게 나타난다는 단점이 지적되었다[8].

입력과 출력간의 대응 관계를 비선형 모델링 (nonlinear modelling)에 의해 표현하는 신경 회로망 (neural network) 역시 음소 경계의 수정에 사용되고 있다[9,10]. 신경망을 사용한 기법에서는 음성 신호로부터 직접적으로 얻어질 수 있는 다양한 특징 변수를 입력으로 하여, 신경망의 출력이 1 또는 0 인가에 따라 음소 경계를 판단하게 된다. 이때 단일 신경망을 사용하여 음소 경계를 예측하는 경우, 음소 전이 특성 (phone transition characteristics)를 반영하지 못하고, 전체적인 음소 경계의 특성에 의해서만 분할을 수행하므로 HMM을 단독으로 사용하는 경우와 비교하여 성능 향상이 뚜렷하지 못하다[9]. 이에 따라 음소 전이의 패턴에 따라 개별적인 신경망을 구성하는 방법이 제안되었다[9,10]. Toledano에 의해 제안된 기법에서는 음소를 유성음소군 (voiced phoneme group) 과 무성음소군 (unvoiced phoneme group)으로 구분하여 총 4개의 음소군 조합에 대해 개별 신경망을 학습시키는 방법이 사용되었다. 실험 결과에 따르면, 단일 신경망을 사용하는 경우와 거의 동일한 결과를 얻은 것으로 보고되어, 음소군에 따른 개별 신경망의 사용이 성능 향상과 직결되지는 않은 것으로 잠정적인 결론을 내렸다.

그러나 보다 세분화된 음소 전이 정보를 사용한 박 등의 방법[10]에서는 단일 신경망을 사용하는 경우보다 향상된 성능을 얻어, Toledano의 연구와는 상반된 결과를 보고하였다.

이와 같은 복수 신경망을 음소 경계의 수정에 이용해도 항시 향상된 성능을 보이지 않는 이유로는 방법[9]가 유/무성음의 구분이라는 비교적 간단한 방법에 의해 음소 전이 패턴을 구분하였다는 점과 두 방법 모두 단지 경험적인 방법에 음소 전이 패턴을 구분하였으며, 경계 오차를 최소화하는 관점은 반영되지 못했다는 것으로 요약할 수 있다.

본 논문에서는 음소 전이 패턴의 분류를 신경망의 학습 과정에서 얻을 수 있는 새로운 신경망 학습 알고리즘을 제안하였다. 제안된 방법은 사용자가 신경망의 개수만을 지정하면, 음소 전이 패턴을 자동적으로 분할하고, 각각의 분류 패턴에 대해 최적의 신경망이 구성되도록 하였다. 여기서 "최적"의 신경망은 추정된 음소 경계와 수작업 레이블링에 의해 얻어진 음소 경계간의 전체 자승 오차가 최소화 되는 관점에서의 "최적"을 의미한다. 제안된 기법의 성능 평가를 위해 신경망에 의해 수정된 음소 경계와 수작업 음소 경계간의 전체적인 오차를 계산하였으며 수동 레이블링 작업의 대치 가능성을 알아보기 위해 전체 음소 중 몇 개의 음소가 음질적인 저하가 일어나지 않는 범위내에서 음소 분할이 이루어지는가를 조사하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서는 본 논문에서 구성한 자동 레이블링 시스템의 전체적인 구조에 대해 설명하며, 3장에서는 최적 신경망을 구성하기 위한 음소 전이의 분할과 신경망 학습 방법이 제시되었다. 4장에서는 실험 결과를 통해 기존 기법과의 성능을 비교하였으며, 마지막으로 5장의 결론으로 본 논문을 끝맺었다.

II. 신경망과 HMM을 이용한 자동 음소 레이블링 시스템

그림 1에 본 논문에서 구성한 자동 레이블링 시스템의 블록도가 제시되었다. 학습 데이터로부터 각 음소에 대한 HMM을 생성하고, 온라인 (on-line) 음소 분할에서는 학습된 HMM과 음소열, 음성 데이터를 이용하여 음소 경계를 얻는다. 음소열을 주어진 음성에 얼라인먼트 (alignment)하는 과정은 주어진 음소열이 해당 음성 신호에 대해 확률적으로 가장 높은 음소 경계를 찾는 과정으로 설명할 수 있으며, 이는 비터비 알고리즘 (Viterbi algorithm)에 의해 구현된다. HMM에 의해 얻어진 음소 경계는 다음 단의 수정 알고리즘 (refinement algorithm)에 의해 수작업에 의해 얻어진 음소 경계와 더욱 가까워지도록 재조정된다. 본 논문에서는 신경망 (Neural Network)의 일종인 다층 퍼셉트론 (Multi-Layer Perceptron; MLP)에 의해 음소 경계를 수정하도록 하였다. 각 단계를 자세히 살펴보면 다음과 같다.

2.1. HMM을 이용한 음소 분할

본 논문에서 사용한 음성 합성기에서는 49개의 단음소 (monophone), 23개의 이중음소 (diphone), 1개의 묵음 (silence)와 1개의 단묵음 (short-pause) 포함하는 총 74개의 음소를 음성 합성의 기본 단위로 사용하였다. 따라서 HMM의 생성을 위한 학습 데이터는 74개의 음소들로 레이블 되어 있다. 본 논문에서 사용된 HMM은 전형적인 좌-우 모델 (left-to-right model)이 사용되었으며, 각 스테이트에 대한 확률 밀도 함수는 혼합 가우시안 함수를

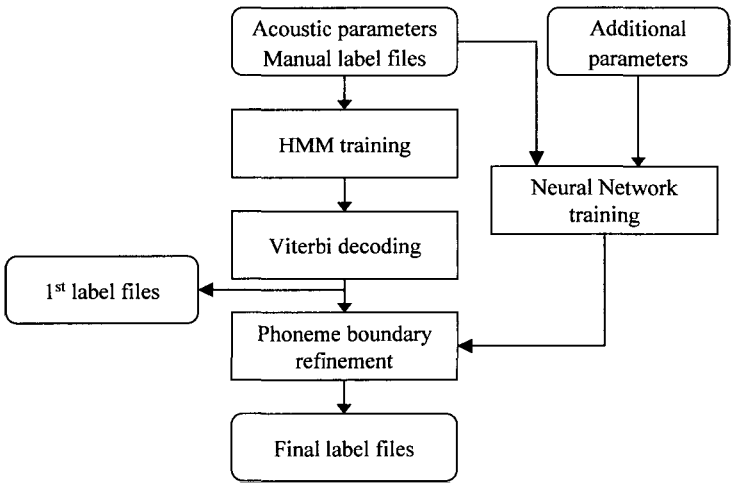

그림 1. 자동 레이블링의 전체 블럭도
Fig. 1. Overall block diagram of automatic labelling.

사용하였다. 즉 연속 관찰 HMM (continuous observation HMM)이 사용되었다. 특징 변수로는 13개의 MFCC (Mel Frequency Cepstrum Coefficient)와 13개의 차분-MFCC (delta MFCC), 13개의 차분-차분-MFCC (delta-delta MFCC)를 포함하는 총 39개의 변수가 사용되었다. 이들 특징 변수는 25 msec의 길이를 갖는 해밍 창함수 (hamming window)를 10 msec마다 이동시켜 가며 계산된다. 따라서 HMM 음소 분할시의 시간 해상도 (time resolution)은 0.01초가 된다.

각 음소에 대한 HMM의 생성, 온라인 과정에서의 음소 얼라인먼트는 MS (MicroSoft)사의 HTK (Hidden Markov Model Tool Kit)을 사용하였다.

HMM을 구성하는 변수인 스테이트수, 혼합 가우시안 모델에서 단일 가우시안 함수의 개수 등은 수차례의 실험을 통해 결정하였다. 몇몇 실험에 의하면, 단일 화자 (speaker)만을 대상으로 하는 음소 레이블링과 같은 응용에서는 각 스테이트의 확률 밀도 함수를 단일 가우시안 함수만을 사용해도 충분하다고 보고하였으나, 본 논문에서 수행한 실험 결과에 따르면 단일 가우시안 함수를 사용한 경우가 혼합 가우시안을 사용한 경우에 비해 성능 저하가 비교적 크게 나타나는 것이 관찰되었다. 실험 결과 모든 음소에 대해 5개의 스테이트를 갖고, 3개의 단일 가우시안 함수를 사용한 경우, 계산량과 음소 레이블링의 정밀도 면에서 우수한 결과를 나타냄을 알 수 있었다.

2.2. MLP를 기반으로 하는 음소 경계의 수정

MLP에 입력되는 특징 변수과 음소 경계 근방에서의 MLP 목표 출력간의 관계가 그림 2에 제시되었다. 그림에서 보듯이 음소의 경계에 해당하는 부분에서는 MLP의 목표 출력이 "1"이 되도록 설정하였으며, 그렇지 않은 부분에서는 MLP의 출력이 "0"이 되도록 하였다. 그러나 이와 같이 음소 경계에서 MLP 출력이 급격하게 변동하는 경우, 음소 경계에서 그릇된 MLP 출력이 얻어질 소지가 있으므로 음소 경계를 기준으로 좌, 우 영역에 대해서는 출력이 0.5가 되도록 하였다[10].

MLP의 입력 변수로는 HMM에서 사용된 13개의 MFCC와 함께, 인접한 두 개의 프레임에서 계산된 스펙트럼 특징변수의 변화율 (Spectral Feature Transition Rate; SFTR)[12], 단구간 영교차율 (Short-Time Zero Crossing Rate; ZCR), 대칭 컬백-라비블러 거리 (Symmetrical Kullback-Leibler Distance; SKLD)[13]가 사용되었다. 이중 SFTR은 스펙트럼의 변화정도를 반영하는 변수로, 음성 신호의 시간축 분할 (temporal decomposition) 기법에 사용되고 있으며[12], 단구간 영교차율은 무성 자음과 유성 모음간의 구분을 용이하게 하는 변수라는 점을 감안하여 포함되었다. SKLD는 Klabber 등의 연구에서 음소 경계면에서의 청각적인 이질감을 가장 잘 표현하는 변수로 알려져 있으며[13], 따라서 청각적인 상이성에 따라 음소 경계를 구분짓는 수동 레이블링과 근접한 결과를 얻는데 유익한 변수로 간주되어 신경망의 특징변수로 이

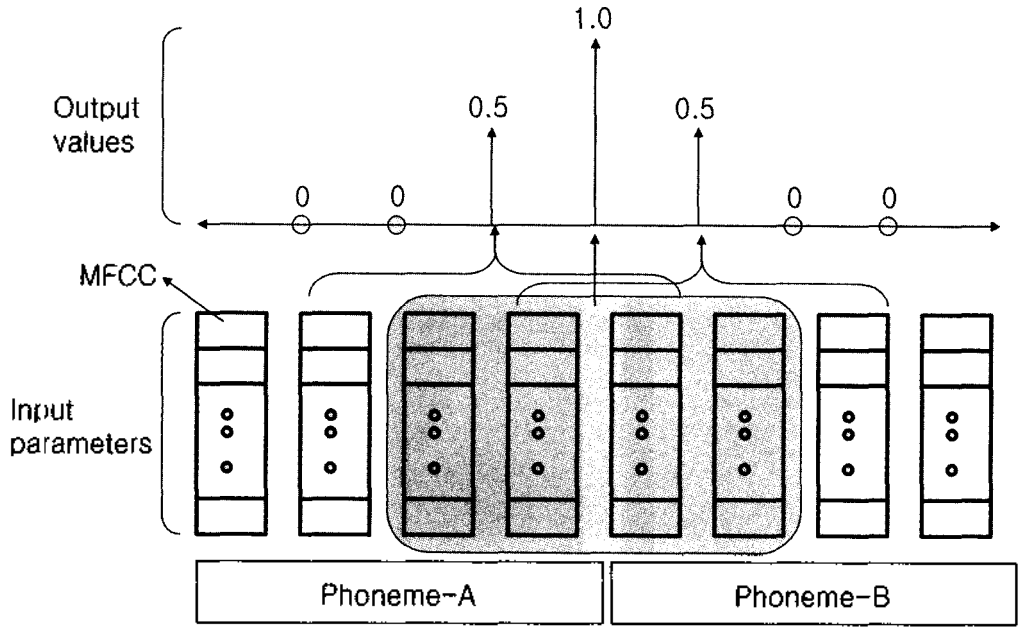

그림 2. MLP의 입력과 목표 출력간의 관계
Fig. 2. Relationship between MLP input and target output.

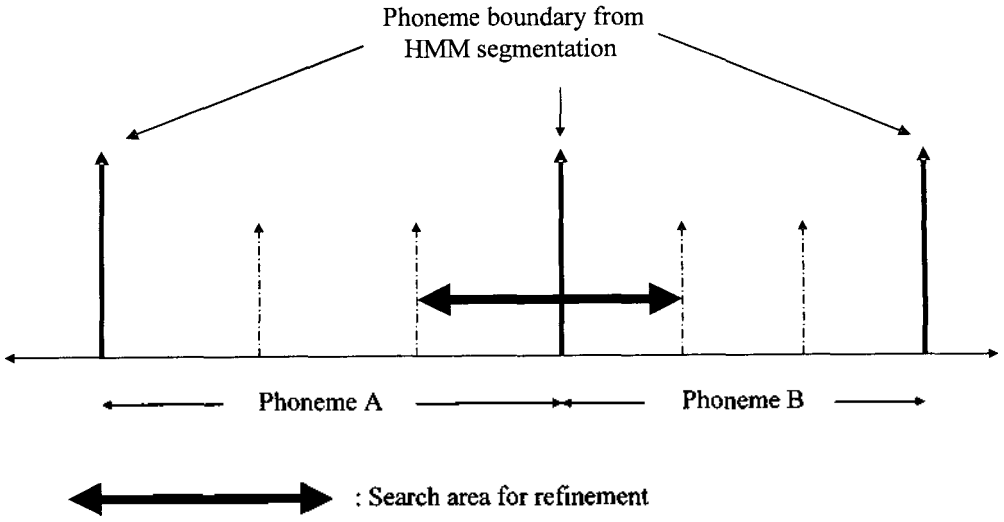

그림 3. 음소 경계 수정 영역
Fig. 3. Phone boundary refinement area.

용하였다.

MLP를 구성하는 요소로서 계층 (layer)의 개수, 각 계층에서의 노드수, 노드에 대한 비선형 함수의 종류등을 들 수 있는데, 이들 변수들은 수학적인 분석에 의해 최적의 값을 찾을 수 없으므로 본 논문에서는 수 차례의 실험을 통해 최적의 값을 찾도록 하였다. 실험적인 결과에 의하면 계층의 수는 1개의 은닉 계층 (hidden layer)을 포함된 경우가 1개 이상의 은닉 계층이 포함된 경우와 비교하여 큰 차이를 나타내지 않았으며, 따라서 계산량을 절감할 수 있는 1개의 은닉 계층을 갖는 MLP가 본 논문에서 사용되었다. 또한 은닉 계층에서의 노드수는 15개로 설정하였는데 이 값 또한 실험을 통해 결정하였다.

신경망의 학습에는 에러 역전파 알고리즘 (error-propagation algorithm)[14]이 사용되었다. 에러 역전파 알고리즘은 MLP를 구성하는 각 가중치를 출력단에서의 에러가 최소화되도록 반복적으로 수정하는 방법이다. 본 논문에서는 음소 경계에서 강조된 에러를 갖도록 출력단에서의 에러 가중치를 적응적으로 조정하였다. 즉 MLP의 목표 출력이 1 (음소 경계)이고, 실제 MLP의 출력이 0.5보다 작은 경우에는 출력단에서의 에러를 2배 증가시키도록 하였다. 이와 같은 출력단에서의 적응적인 에러 강조는 음소 경계에서의 오차에 대해 민감하게 반응하도록 MLP를 학습시킬 수 있으며, MLP의 출력이 "0"인 경우가 출력이 "1"인 경우와 비교하여 매우 많이 발생하는 편향된 학습 데이터 (biased-training data) 문제를 해결할 수 있는 장점을 갖는다.

온라인 음소 경계의 수정 시에는 MLP의 학습 과정에 사용했던 동일한 입력 변수를 MLP에 연속적으로 입력하고, 그 출력이 1에 가까운 지점을 수정된 음소 경계로 간주하였다. 이때 1에 가까운 출력을 검색하는 구간이 너무 큰 경우에는 MLP가 이상출력 (outlier)을 발생하는 경우, 수정된 음소 경계 위치가 HMM으로 추정된 음소 경계보다 오히려 더 큰 오차를 나타낼 수 있다. 반대로 추정 구간이 너무 작은 경우에는 수정된 음소 경계가 HMM 음소 경계에 지나치게 의존적이라는 문제점이 발생한다. 본 논문에서는 이와 같은 문제를 해결하기 위하여 그림 3에서와 같이 HMM에 의해 1차적으로 추정된 음소 경계를 기준으로, 좌, 우 음소 길이의 1/3 되는 길이 만큼을 좌, 우 이동시켜 가면서, MLP의 출력을 구하고, 이 출력값이 최대가 되는 지점을 수정된 음소 경계로 간주하였다. 이와 같은 제한된 탐색 방법은 전영역 탐색 방법에 비해 계산량을 줄일 수 있으며, 앞서 언급한 바와 같이 MLP의 이상 출력으로 인한 영향을 억제할 수 있는 장점을 갖는다.

III. 최적 신경망 구성을 위한 음소 전이 패턴의 분류

음소가 전이 (transition)되는 부분에서의 스펙트럼 특성은 전이되는 음소의 패턴에 따라 다르게 나타나는 것으로 가정할 수 있다. 즉 무성 자음에서 유성 모음으로 전이되는 영역에서는 스펙트럼의 불연속성이 강하게 나타나며, 이에 따라 MFCC의 변화량, SFTR같은 변수도 큰 값을 갖게 된다. 이와 달리, 유성 모음과 유성 자음 (종성) 등으로 전이되는 영역에서는 완만하게 변화하는 스펙트럼이 관찰된다. 즉 음소 경계를 구분하는 알고리즘은 전이되

는 음소의 패턴에 따라 적응적으로 결정되어야만 우수한 성능을 나타낼 수 있다.

이를 반영한 기법이 Toledano와 Park 등에 의해 제안되었으며, 이 두 기법은 모두 음소 전이 패턴을 경험적인 방법에 의해 결정하였다. 이러한 경험적인 분류 기법은 대상 언어에 대한 사전 지식이 필요하며, 음소 전이 패턴의 분류와 신경망의 학습이 독립적으로 이루어지므로 획득된 신경망이 학습 데이터에 대해 최적의 신경망을 보장하지는 못한다고 볼 수 있다. 따라서 본 논문에서는 언어에 대한 사전 지식 없이, 최소 추정 오차 면에서 최적의 성능을 나타낼 수 있는 분류 신경망을 자동적으로 구성하는 알고리즘을 제안하였다.

음소의 전이 패턴에 따라 신경망을 구분하여 학습시키는 문제는 다음과 같은 최소화 문제 (minimization problem)으로 정의할 수 있다.

$$\Phi^* = \arg\min_{\Phi}\Big[\sum_{k=1}^{K}\sum_{P_j\in s_k}\min_{k}\Big\{\sum_{P_c(m)=j}\sum_{n\in\Delta_m}|y_d(n) - F(\Phi_k, X(n))|^2\Big\}\Big] \quad (1)$$

윗 식에서 각 변수가 의미하는 값은 다음과 같다.

K : 전체 신경망의 개수 (분류 패턴의 개수).

$y_d(n)$: n번째 프레임에서의 목표 출력값 (음소 경계=1)

$F(\Phi_k, X(n))$: 입력 신호 $X(n)$에 대한 k번째 신경망 Φ_k의 출력.

P_j : j번째 음소 조합의 인덱스.

s_k : k번째 분류 패턴의 집합.

$P_c(m)$: m번째 음소 경계에 대한 음소 조합 인덱스.

한편 Δ_m은 m번째 음소 경계에 대한 주변 프레임 인덱스 값들을 의미하는데, 이를 수식으로 나타내면 다음과 같다.

$$\Delta_m = \{t \mid \frac{t_{m-1}+t_m}{2} \le t \le \frac{t_m+t_{m+1}}{2}\} \quad (2)$$

여기서 t_m은 m번째 음소 경계의 프레임 인덱스를 나타낸다.

위와 같은 최소화 문제를 해결하기 위해, 본 논문에서는 다음과 같은 반복 추정 알고리즘을 제안하였다. 제안된 알고리즘은 그림 4에 제시된 바와 같이, 초기 신경망 집합을 이용하여 학습 데이터를 오차가 가장 작도록 분류하며, 동일하게 분류된 학습 데이터들만으로 새롭게 신경망을 학습시킨다. 여기서 구성된 신경망 집합을 다시 분류에 사용하여 이러한 과정을 반복적으로 수행하여 최종적인 신경망 집합을 구성한다. 알고리즘을 단계별로 자세히 살펴보면 다음과 같다.

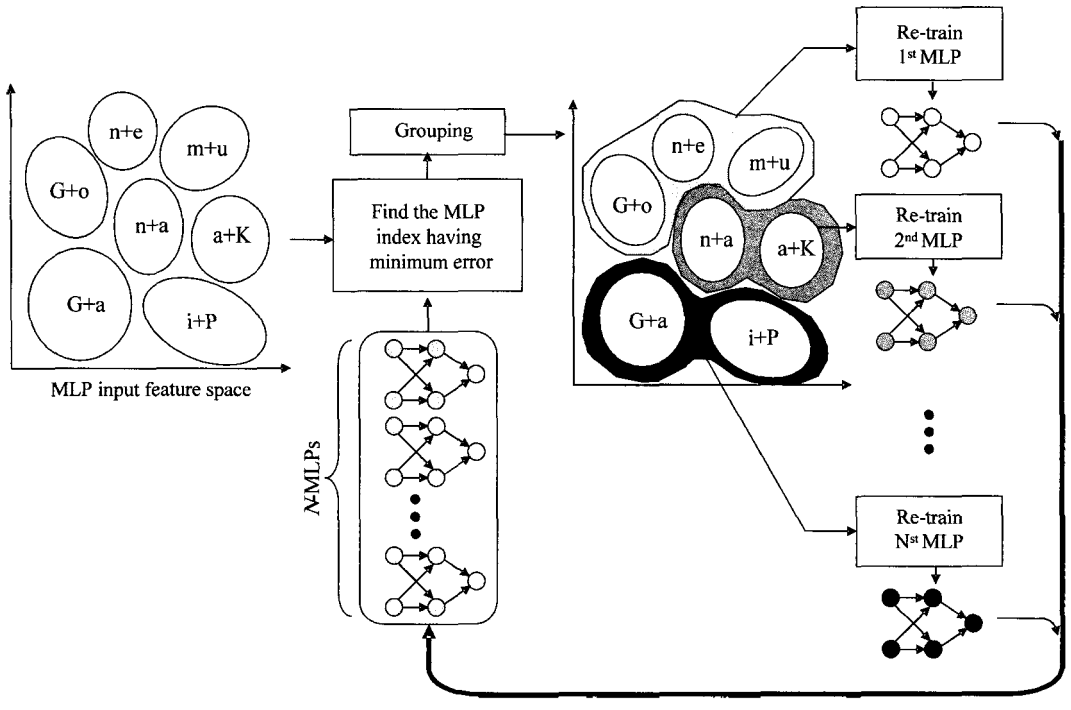

그림 4. MLP set의 생성 과정
Fig. 4. MLP set designing procedure.

단계 1) 초기화; 학습 데이터쌍 $\{X(n), y_d(n)\}_{n=1}^{N}$ 과 초기 신경망 집합 $C_0=\{\Phi_k^0\}_{k=1}^{K}$ 을 구성한다. 또한 각 음소 경계에 대한 시간 $\{t_m\}_{m=1}^{M}$ 을 수동 레이블링 작업을 통해 준비하고 음소 경계에서의 음소 전이 패턴 $P_c(m)$을 구한다. 음소 전이 패턴은 좌/우 음소의 모든 조합중의 하나로 표현된다. 문턱치 ϵ을 적절한 값으로 설정하고 초기 왜곡 $D_{-1}=\infty$, 반복 인덱스 $i=0$으로 설정한다.

단계 2) 분류; 음소 전이 패턴 $P_c(m)$이 동일한 모든 학습 데이터를 동일한 그룹으로 간주하여, 모든 음소 전이 패턴에 대해 최적의 신경망 인덱스를 구한다. 최적의 신경망 인덱스는 아래와 같이 K개의 신경망 중에서 목표 출력값과 실제 출력값과의 전체 자승 오차가 최소화되는 신경망을 택함으로서 얻어진다.

$$c_i(P_j)=\arg\min_k\{\sum_{P_c(m)=j}\sum_{n\in\Delta_m}|y_d(n)-F(\Phi_k^i, X(n)|^2\} \tag{3}$$

여기서 $c_i(P_j)$는 j번째 음소 조합에 대한 i번째 반복 학습에서의 최적 신경망 인덱스를 나타낸다.

단계 3) 학습 데이터의 재배열; 모든 학습 데이터들을 단계 2)의 분류에 따라 얻어진 신경망 인덱스에 따라 재구성한다. 재배열된 학습데이타는 K개의 셀 (cell)로 이루어진 공간 (space)으로 표현된다. 각 셀을 s_k^i로 표현한다면 재배열된 학습 데이터가 표현되는 공간은 아래와 같이 집합 형태로 표현될 수 있다.

$$A^i=s_1^i\cup s_2^i\cup,\cdots,\cup s_K^i \tag{4}$$

이때 각 셀 s_k^i에 포함되는 모든 학습 데이터들은 동일한 신경망을 최적의 신경망으로 갖는다. 따라서 셀 s_k^i는 아래와 같이 표현될 수 있다.

$$s_k^i=\{\{X(n), y_d(n)\}|n\in\Delta_m,\ \ where\ P_c(m)\in W_k^i\} \tag{5}$$

여기서

$$W_k^i=\{P_j|c_i(P_j)=k\} \tag{6}$$

단계 4) 수렴 확인; 재배열된 학습 데이터와 각 셀에 대한 신경망을 이용하여 현재 반복 단계에서의 전체 왜곡을 구한다.

$$D_i=\sum_{k=1}^{K}\sum_{P_j\in s_k}\min_k\{\sum_{P_c(m)=j}\sum_{n\in\Delta_m}|y_d(n)-F(\Phi_k, X(n))|^2\} \tag{7}$$

여기서 구한 왜곡값과 이전의 왜곡값간의 변화율을 구하여 이 변화율을 문턱치와 비교한다. 즉 $(D_{i-1}-D_i)/D_i\le\epsilon$ 이면 학습 과정을 중단하고 단계 3에서 얻어진 셀 A^i, W_k^i 와 현재의 신경망 집합 $C_i=\{\Phi_k^i\}_{k=1}^{K}$을 최종적으로 얻어진 음소 전이 패턴의 분할과 해당 패턴에 대한 최적으로 신경망으로 사용한다. 그렇지 않은 경우에는 다음 단계를 수행한다.

단계 5) 신경망의 재학습; 동일한 셀에 포함되는 학습 데이터들을 이용하여 해당 셀의 신경망을 재학습한다. 신경망의 재학습은 에러 역전파 알고리즘이 사용되며, k번째 신경망에 대한 신경망 파라메터의 갱신식은 아래와 같이 주어진다.

$$\Delta\Phi_k^i=\eta\sum_{\{X(n),y_d(n)\}\in s_k^i}\nabla_{\Phi_k^i}\frac{1}{2}|y_d(n)-F(\Phi_k^i, X(n))|^2 \tag{8}$$

윗 식은 k번째 신경망의 갱신이 k번재 셀에 포함되는 학습 데이터에 의해서만 이루어짐을 의미한다. K개의 신경망에 대해서 학습이 완료되면, 새롭게 구성된 신경망들로 신경망 집합 C_i를 구성한 뒤, 반복 횟수 i를 1증가시키고 단계 2)로 이동, 위의 과정을 반복적으로 수행한다.

수렴성이 확인되어 분류/재학습이 완료되면, 온라인 과정에서는 단계 3에서 얻어지는 각 셀에 대한 음소 조합의 패턴들과 셀에 대한 최적의 신경망을 이용하여 음소 경계를 수정한다.

IV. 실험 및 결과

제안된 기법의 검증을 위한 음성 데이터로, 사용된 음성 합성기의 데이터 베이스로부터 1,000개의 문장을 사용하였다. 이 음성 데이터는 55,250개의 음소 경계를 포함하며, 신경망의 학습에는 총 476,902개의 학습 데이터가 사용되었다. 학습 데이터에 포함되지 않은 문장에 대한 성능 평가를 위해 1,000개의 문장은 500개의 문장을 갖는 두 개의 세트로 구분되어 각 세트는 신경망의 학습, 테스트에 사용되었다. 테스트 데이터들은 다음의 4가지 방법에 의해 음소 분할한 후, 각 방법간의 비교를 통해 제안 방법의 성능을 평가하였다.

1) HMM 만을 단독으로 사용하여 음소 분할
2) HMM을 사용하여 음소 분할된 결과를 단일 신경망을 통해 수정하는 경우
3) HMM을 사용하여 음소 분할된 결과를 좌, 우 음소의 유/무성음 특성에 따라 구분된 4개의 신경망으로 수정하는 경우[9]
4) HMM을 사용하여 음소 분할된 결과를 제안된 분류/재학습 기법을 통해 얻어진 4개의 신경망으로 수정하는 경우

성능 평가를 위한 척도로는 본 논문의 주된 목적이 자동 레이블링된 결과가 수동 레이블링의 결과와 되도록 유사하도록 하는 것이므로, 자동 레이블링의 음소 경계와 수동 레이블링 음소 경계간의 제곱근 평균 자승 오차 (Root Mean Square Error; RMSE), 평균 절대 오차 (Mean Absolute Error; MAE)를 사용하였다. 또한, 수동 레이블과 비교하여 20 msec 이내의 경계 오차를 갖는 자동 레이블링된 음소에 대해서는 합성음의 품질이 크게 저하되지 않는다는 실험적인 경험을 바탕으로, 20 msec 이내의 경계 오차를 갖는 음소 경계가 전체 음소 경계중 몇 %를 차지하는가를 조사하였다.

이에 대한 결과를 표 1에 제시하였다. 표에서 보면 신경망을 후처리기로 사용한 모든 경우에 있어서, HMM을 단독으로 사용하는 경우보다 성능 향상이 얻어짐을 알 수 있다. 이는 신경망이 음소 경계의 검출에 유용하게 사용될 수 있음을 의미하는 것이다. 단일 신경망의 사용과 복수 신경망을 사용하는 경우와의 비교는 복수 신경망의 사용이 더욱 향상된 결과를 얻는 것으로 관찰되었으며, 신경망의 학습과정에서 음소 전이 패턴을 분류하고 적절한 신경망을 구성하는 제안된 기법이 경험적인 방법에 의해 복수 신경망을 학습시키는 방법보다 근소하게 우수한 성능을 나타내고 있다. 두 기법간의 성능 비교에서 RMSE와 MAE는 근소한 차이를 보이지만, 20 msec 이내의 경계 오차를 갖는 음소의 비율에 있어서는 제안된 기법이 2.2% 높은 값을 갖음을 알 수 있다. 이를 사용된 전체 음소 갯수로 환산하면 1215개의 음소에 해당되며, 따라서 음질적인 저하를 일으킬 수 있는

표 1. 각 방법에 따른 자동 레이블링의 성능 비교
Table 1. Performance of automatic labeling.

방법	RMSE (msec)	MAE (msec)	% error<20 msec
HMM only	13.5	9.3	90.4
HMM+single MLP	12.2	7.7	91.2
HMM+4 MLPs (off-line)	10.7	6.8	93.0
HMM+4 MLPs (proposed)	10.1	6.2	95.2

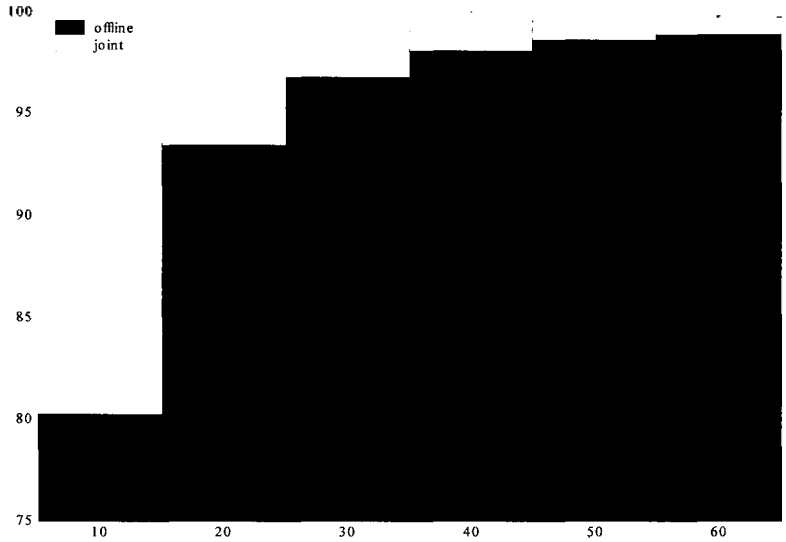

그림 5. 유/무성음 정보로 구분된 다중 MLP를 사용하는 경우와 제안된 기법에 의해 생성된 다중 MLP를 사용하는 경우의 음소 경계 위치에 대한 누적 오차
Fig. 5. Cumulative distribution of the difference in location of the phone boundaries obtained by the proposed method and the refinement algorithm with multiple MLPs specialized by the voicing status of the phoneme.

음소의 빈도가 제안된 기법을 통해 상당수 감소되었음을 의미한다.

그림 5는 방법 3)과 방법 4)에 대한 누적 경계 오차 분포를 나타낸다. 즉, 가로축의 10 msec에 대응되는 세로축의 80% 또는 83%는 10 msec 이내의 경계 오차를 갖는 음소들의 전체 음소 개수에 대한 비율을 나타낸다. 누적 경계 오차의 분포는 가로축의 값이 증가함에 따라 100% 지점에 수렴되는 특성을 갖게 된다. 그림에서 관찰되는 바와 같이, 누적 경계 오차면에서도 제안된 기법에 의해 생성된 음소 경계가 기존의 경험적인 분류/학습 신경망 기법에 비해 우수한 성능을 나타냄을 알 수 있다.

이와 같은 객관적인 척도에 의한 결과를 고려하면 제안된 기법은 수동 레이블링된 음소 경계와 매우 유사한 음소 경계를 추정하는 것으로 판단되며, 일정량의 학습 데이터가 확보된 상태에서 많은 시간이 소요되었던 수동 레이블링 작업을 학습된 HMM과 신경망에 의해 자동화된 기법으로 대치할 수 있을 것으로 생각된다.

V. 결 론

본 논문에서는 음성 합성기의 방대한 데이터 베이스 구성에 유용하게 이용될 수 있는 자동 음소 분할의 한 기법을 제안하고 성능을 평가하였다. 제안된 기법은 은닉 마코프 모델을 이용한 통계적인 분할 방법에 따라 1차적으로 음소 경계를 추정하고, 여기서 얻어진 음소 경계를 좌, 우로 미소하게 이동시켜 수동 레이블링에 의한 음소 경계와 더욱 가까운 음소 경계를 얻도록 하였다. 음소 경계의 수정에는 비선형 대응관계를 표현하는 신경 회로망이 이용되었으며, 음소 전이 패턴에 따라 적응적인 신경망이 사용될 수 있는 기법이 제안되었다. 전체 음소 전이 패턴의 분류와 각 분류 패턴에 대한 최적의 신경망은 분류와 학습이 유기적으로 결합된 학습 알고리즘에 의해 구현되었으며, 이 알고리즘은 기존의 경험적인 분류에 의한 신경망의 학습 기법에 비해 실험상으로 우수한 성능을 나타내었다.

본 논문에서 제시된 방법을 통해 얻은 음소 경계의 정밀도는 20 msec 이내의 경계 오차를 갖는 음소의 개수가 전체 음소중 95%를 상회하는 것으로 나타났으며, 이는 데이터 베이스의 작성시 자동 레이블링 기법이 적용되더라도 수동 레이블링시와 필적하는 결과를 얻을 수 있음을 의미한다. 따라서 본 논문에서 제안된 기법은 개발 시간의 지연, 비용 상승, 음소 경계의 일관성 부족이라는 수동 레이블링의 문제점을 효과적으로 해결해 줄 수 있을 것으로 판단된다.

감사의 글

본 연구는 삼성전자(주) 종합기술원의 연구비 지원에 의한 결과임.

참 고 문 헌

1. Y. Sagisaka, "Speech synthesis from text," *IEEE Communications Magazine*, 28 (1), 35–41, January, 1990.
2. A. J. Hunt, and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP '96*, 1, 373–376, 1996.
3. Y., Stylianou, T. Dutoit, and J. Schroeter, "Diphone concatenation using a harmonic plus noise model of speech," *Proc. EUROSPEECH '97*, 613–616, 1997.
4. A. J. Hunt and A. W. Black, "Concatenative speech synthesis using units selected from a large speech database," *Draft paper*, 1997.
5. M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou and A. Syrdal, "The AT\&T Next-Gen TTS system," *Proc. Joint Meeting of ASA, EAA, and DAGA*, Berlin, Germany, March 1999.
6. Jan P. van Hermert, "Automatic segmentation of speech," *IEEE Trans. Signal Processing*, 39 (4), 1008–1012, 1991.
7. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and T. M. Zampini, "A bootstaping training technique for obtaining demisyllable reference patterns," *Journal of Acoust. Soc. Amer.*, 71, 1588–1595, 1982.
8. A. Bonafonte, A. Nogueiras and A. R.-Garrido, "Explicit segmentation of speech usnig gaussian models," *Proc. IEEE Int. Conf. Spoken Language Processing*, 1269–1272, 1996.
9. D. T. Toledano, "Neural network boundary refining for automatic speech segmentation," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 3438–3441, 2000.
10. E.-Y. Park, S.-H. Kim and J.-H. Chung, "Automatic speech synthesis unit generation with MLP based postprocessor against auto-segmented phoneme errors," *Proc. International Joint Conference on Neural Networks*, 2985–2990, 1999.
11. L. Wu, M. Niranjan, and F. Fallside, "Nonlinear predictive vector quantization with recurrent neural nets," *Proc. IEEE -SP Workshop on Neural Networks for Signal Processing*, 372–381; Baltim ore, MA, 1993.
12. A. C. R. Nandasena and M. Akagi, "Spectral stability based event localizing temporal decomposition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 3438–3441, 1998.
13. E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. on Speech and Audio Singal Processing*, 9 (1), 39–51, 2001.
14. R. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, 4–22, April, 1987.

저자 약력

이 기 승 (Ki-Seung Lee)

1991년 2월: 연세대학교 전자공학과 (공학사)
1993년 2월: 연세대학교 대학원 전자공학과 (공학석사)
1997년 2월: 연세대학교 대학원 전자공학과 (공학박사)
1997년 3월~1997년 9월: 연세대학교 신호처리 연구센터 선임 연구원
1997년 10월~1999년 8월: AT&T Shannon Lab, Consultant
1999년 9월~2000년 9월: AT&T Shannon Lab, Senior Technical Staff Member
2000년 11월~2001년 8월: 삼성종합기술원 HCI Lab 전문연구원
2001년 9월~현재: 건국대학교 정보통신대학 전자공학부 조교수

※ 주관심분야: 음성 합성, 운율 제어, 음성 변환, 음성 부호화기 등

김 정 수 (Jeong-Su Kim)

1988년 2월: 연세대학교 전산과학과 (이학사)
1990년 2월: 한국과학기술원 전산학과 (공학석사)
1990년 3월~1993년 1월: 삼성전자 정보통신연구소 연구원
1993년 2월~현재: 삼성종합기술원 HCI Lab 전문연구원

※ 주관심분야: 음성합성, 대화 에이전트, 자연어처리