남성 음성 triphone DB 구축에 관한 연구

Dialogic Male Voice Triphone DB Construction

김 유 진*, 백 상 훈*, 한 민 수**, 정 재 호*
Yu Jin Kim*, Sang Hoon Baek*, MinSoo Han**, Jae Ho Chung*

요 약

본 논문에서는 음성합성을 위한 대화체(회화체) 음성의 triphone 단위[1] 데이타베이스 구축에 대하여 보고한다. 특히 본 연구는 방송 매체를 이용하여 대화체 음성을 수집하고 3차에 걸친 대화체 표기(transcription)작업을 거쳐 triphone 단위의 분할 및 음성기호층 단계의 레이블링을 목표로 진행되었다. 수집된 총 10시간 방송분량중 6시간 분량을 데이타베이스 구축에 사용하였으며, 나머지 4시간은 예비 분으로 수집되었다.

낭독체 음성 데이타베이스 구축과는 여러 면에서 다른, 대화체 음성 데이타베이스 구축을 위한 음성 데이타 수집에서부터 triphone 단위 레이블링까지의 과정을 본 논문에서 기술하고, 보다 체계적이고 일관성있는 대화체 음성 데이타베이스 구축을 위해 필요한 계획 및 요구 사항에 대해서 논하고자 한다.

Abstract

In this paper, dialogic triphone data base construction for triphone synthesis system is discussed. Particularly, in this work, dialogic speech data is collected from the broadcast media, and three different transcription steps are taken. Total 10 hours of speech data are collected. Among them, six hours of speech data are used for the triphone data base construction, and the rest four hours of data are reserved.

Dialogic speech data base construction is far different from the reciting speech data base construction. This paper describes various steps that necessary for the dialogic triphone data base construction from collecting speech data to triphone unit labeling.

Ⅰ. 서 론

최근 음성 신호 처리 기술을 적용한 응용 범위는 날로 넓어져 가고 있으며 그에 따라 관련 연구도 활발히 진행되고 있다. 그러나 아직도 한국어 음성을 이용한 음성 관련 연구의 발전에 걸림돌이 되는 것은, 합성/인식 분야 등의 연구 및 구현된 시스템의 객관적인 평가에 필요한 공용의 음성 데이타베이스가 거의 전무하다는 것이다[1]. 또한 각 연구소나 대학에서 소규모로 구축하여 사용되어 온 기존의 데이타베이스들은 대부분 방음 환경, 사무실 환경, 전화선을 이용한 낭독체 음성 데이타베이스이며, 레이블링 단위도 PBW(Phoneme Balanced Word), 숫자음, 기능어 등의 음절(syllables) 단위의 레이블링 데이타베이스가 주류를 이루고 있다. 하지만 궁극적인 음성의 인식/합성을 위해서는 연속어 처리에 대한 연구 및 연속어 처리를 고려한 대화체 음성 데이타베이스 구축이 시급하다[2, 3, 4, 5].

따라서 본 연구에서는 남성 대화체 음성 데이타베이스의 구축에 관하여 연구하였으며, 합성 시스템 성능 중 가장 중요한 요소인 자연성을 만족하기 위하여, 수집한 대화체 음성을 triphone 단위로 레이블링 하였다. 음성 합성 시스템에서 triphone 단위의 사용이 바람직한 이유는

*인하대학교 전자공학과
** 한국전자통신연구소 음향통신 연구실
접수일자: 1996년 2월 21일

1. 본 논문에서 사용된 'triphone'단위는 합성을 위한 실제 음소로 구성된 '3연속 음소열'이라는 점에서 개념적인 정의인 인식의 'triphone' 단위와는 다르다.

다. 본 연구에서는 인접 음성음과의 상관관계 속에서 그 자신의 특성을 규정 지을 수 있으며, 이는 곧 환경적 효과(contextual effect)를 생각하지 않은 단독 음절보다는 환경적 효과를 명확한 경우가 현실적으로 더욱 능동적인 작업을 수행할 수 있기 때문이다[6].

본 고에서는 이상과 같은 목표로 진행된 연구의 과정 및 보완해야 할 점을 순서적으로 기술하고, 연구 결과를 제시한 후, 차후 데이타베이스 구축에서 대처해야 할 몇 가지 다이고 일관성 있는 제약 및 기술적인 요구 사항에 대해서 논하고자 한다. 아직 국내에서는 방송 매체를 이용한 triphone 단위의 데이타베이스가 구축된 예가 거의 없음을 고려할 때, 본 논문에서 기술하는 초기 화자 선정, 데이타 수집 및 triphone 추출 그리고 구축된 데이타베이스에 대한 통계 처리 결과는 앞으로 시도될 많은 데이타베이스들의 구축 과정에서 동반될 여러 가지의 시행착오를 줄일 수 있으리라 사료된다.

그림 1은 본 고에서 설명할 전체 연구 과정에 대한 블록 다이어그램이다. 전체 과정은 화자 선정, 음성 데이타 수집, transcription 작성, triphone 추출, DB 구축 및 통계 처리 등을 포함한다.

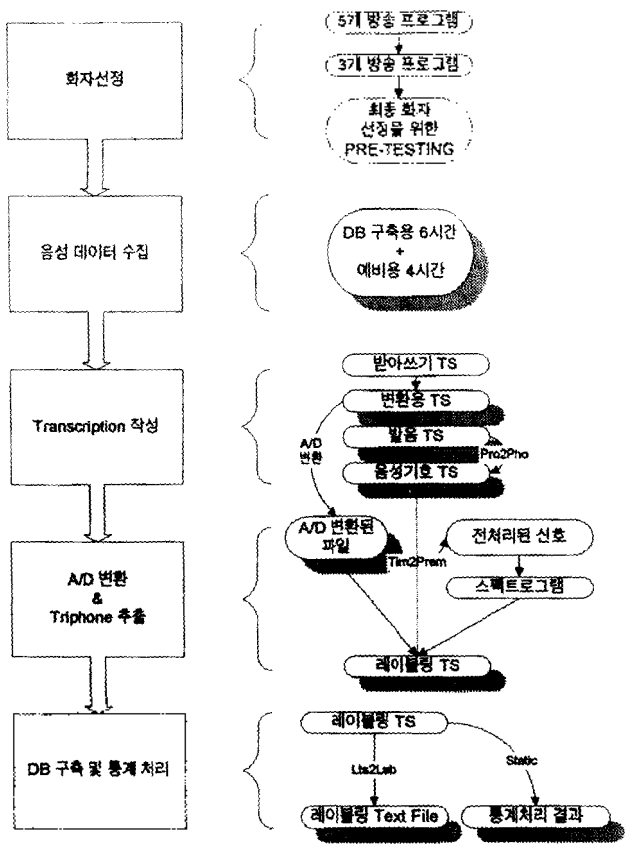


그림 1. 대화체 남성 음성 triphone 데이타베이스 구축 블록 다이어그램

II. 화자 선정

2.1 화자 선정을 위한 사전 작업

본 연구에서는, 표준어와 명확한 발음을 구사하는 화자가, 시 잡음 환경에서 자연스럽게 발성한 일반 대화체의 음성 데이타베이스를 구축하고자 한다. 이러한 기준을 만족하기 위하여 많은 음성 데이타 수집 방안이 제시될 수 있다[3]. 하지만 한정된 연구비, 인력 그리고 시간을 고려하여 최대한 객관적인 만족을 얻을 수 있는 방법으로서 방송 매체를 통한 음성 데이타의 수집 방안을 선택하였다. 따라서, 본 연구에서는 남성 화자가 진행하는 라디오 및 TV 방송의 프로그램을 이용하여 음성 데이타를 수집하기로 한다.

이러한 결정에 따라 사전시험에 의한 프로그램 및 화자의 선정 작업이 수행되었다. 이를 위하여, 우선 3개의 후보 프로그램을 선정하여 1시간 분량씩 수집한 후, 진행자인 남성 화자의 음성에서 적절 소량의 triphone을 추출하여, 단순 접속한 문장을 들어봄으로써 객관적인 선정이 이루어지도록 하는 것이 목표이다. 자세한 내용은 다음과 같다.

우선 수행 연구원들의 시청/청취 의견과 간단한 설문 조사를 통해, 초기 5개의 프로그램이 선정되었으며, 선정된 프로그램들은 다음과 같다.

1) 이상벽, '아침마당', KBS1 TV
2) 박용호, '6시 내고향', KBS1 TV
3) 한선교, '아침 만들기', MBC TV
4) 유 열, '유열의 음악앨범', KBS 라디오
5) 신동호, '신동호가 여는 아침', MBC 라디오

이상 5개의 프로그램은 안정된 진행 방식과 발성 속도, 명확한 발음의 구사 정도 및 합성에 적합한 음색을 기준으로 선정되었다. 또한 선정 과정에서 AM 방송 프로그램은 잡음이 심하다는 매체의 제약과 전체적인 매체의 방송 분위기가 적합하지 않다는 판단에 따라 제외되었다.

초기 선정된 5개의 프로그램들로부터 프로그램의 방송을 1시간씩 수집하였다. 이는 차후 청취 테스트를 통해 triphone 접속 실험에 적용될 3개의 프로그램을 다시 선정하기 위한 사전 작업이다. 수집된 음성신호들 중, (1)번에서 (3)번까지의 TV 프로그램은 오디오 신호만을 DAT로 녹음하였고, (4)번 라디오 프로그램의 경우 테잎을 통해 최대한 저 잡음 환경에서 녹음하였다. 특히, (5)번 라디오 프로그램의 경우는 MBC 방송국에 의뢰하여 DAT로 녹음된 방송을 직접 구입하였다.

추후 실제 데이타베이스 구축 과정을 고려할 때 가장 이상적인 음성 데이타 수집 방법은 (5)번 프로그램의 경우라고 할 수 있다. 이는 데이타 수집이 용이하다는 점 외에 바디오 테잎 또는 카세트 테잎의 경우처럼 A/D 변환과정을 위해 DAT로 옮기는 중간 과정이 필요 없기 때

문이다. 즉 A/D 변환은 원하는 방송 부분에 대해서 선택적으로 이루어지므로 비디오 테잎이나 카세트 테잎보다는 정확한 조작이 가능한 DAT가 편리하다.

참고로, MBC 라디오의 경우 1995년 4월부터 모든 방송 프로그램을 수개월 이상 보존하여 청취자가 원할 경우 그 보존된 내용을 DAT, 카세트 테잎 등의 다양한 매체에 옮겨 판매하고 있다. (DAT인 경우 13,000원/tape.) 그러나, KBS 라디오의 경우 이러한 일반적인 관리 체계가 전혀 세워져 있지 않음을 알 수 있었다. 또한 TV 프로그램의 경우 KBS, MBC 양 방송사 모두 원하는 날짜의 방송 프로그램을 비디오 테잎에 수록하여 판매하고 있었다.

수집된 프로그램의 청취 테스트 과정에서 (1)번과 (5)번 화자들은 최종 화자의 고려 대상에서 제외시켰다. (1)번 화자는 음색이 정교히 다듬어지지 않은 느낌 때문이었다. 이는 (1)번 화자가 원래 아나운서 출신이 아니고, 기자 출신임을 고려할 때, 이에 대한 이유를 찾을 수 있었다. 반면에, (5)번 화자의 경우, 음색은 매우 좋았으나 말의 진행 속도가 너무 빨라 추후 triphone 추출 및 실제 합성 시스템에 이용되었을 경우 많은 어려움이 예상되어 제외되었다. 따라서, triphone 접속 실험을 위하여 선정된 3개의 프로그램은 다음과 같다.

1) 박용호, '6시 내고향', KBS1 TV
2) 한선교, '아침 만들기', MBC TV
3) 유 열, '유열의 음악앨범', KBS 라디오

이러한 선택 과정을 통해 일반적으로 라디오 프로그램의 진행 화자들이 TV 프로그램의 진행 화자들에 비하여 발성 속도가 매우 빠르다는 것을 알 수 있었고, 새벽 방송 시간대 또는 심야 시간대의 방송보다는 오전 8시와 12시 사이 그리고 오후 6시에서 8시 사이의 방송 프로그램들이 비교적 안정된 분위기로 진행된다고 판단되었다. 선정된 3개의 프로그램인 '6시 내고향', '유열의 음악앨범' 그리고 '아침만들기'는 각각 저녁 6시부터 7시, 오전 9시부터 11시, 그리고 오전 8시부터 9시까지 방송되는 프로그램들이었다.

2.2 Triphone 접속

선택된 화자의 triphone들이 실제 합성 시스템에 이용되었을 경우의 결과를 들어보기 위해, 약 10 음절 정도의 문장들을 구성하기로 하였다. 먼저, 1시간 분량의 방송 분량에서 발생 가능한 triphone을 예측하여 문장을 구성한다는 것은 매우 어려운 일이었다. 따라서, 우선 transcription을 작성하여 발생되는 triphone들을 정리하고, 이를 통하여 조합 가능한 의미 있는 문장을 구성하기로 하였다.

Transcription의 작성은 1차와 2차로 나누어 진행되었다. 1차 transcription은 화자가 전하고자 했던 문장을 중심으로 작성되었으며, 2차 transcription은 화자의 발음을 중심으로 진행되었다. 먼저, 2개 TV 프로그램의 경우, 직접 DAT에 녹음된 음성 데이타를 재생하여 들으면서 2명의 화자에 대한 1차 transcription을 작성하였고, 라디오 프로그램의 경우 카세트 테잎에서 화자의 발성 부분을 먼저 샘플링하여 PC에 파일로 저장한 후, 샘플링 파일을 들으면서 1차 transcription을 작성하였다. 이는, 사용된 DAT 플레이어의 경우 전체 시간이 표시되어 있으므로 추후 추출하고자 하는 triphone을 transcription에 기록된 시간으로부터 손쉽게 찾을 수 있지만, 카세트 테잎의 경우 정확한 위치를 찾기가 어렵기 때문이었다. 2차 transcription은 1차 transcription을 바탕으로, 녹음된 음성 데이타를 다시 들으면서 작성하였다. 2차 transcription의 작성 과정에서는, 휴지음, 잡음, 기침 소리 등, 화자가 전하고자 하는 내용 이외의 것은 고려하지 않았다. 또한 triphone 추출 시에 부적합한 외국어의 발음, 불분명한 발음 부분 등도 2차 transcription에서 제외되었다. 위와 같은 과정을 통하여, 결과적으로 구성된 각 프로그램의 합성 문장들은 다음과 같다.

1) 박용호, 6시 내고향
'잔머리가 마니 도라서 어따주리'
2) 유 열, 유열의 음악앨범
'다가서는 절 그대는 보고 인나요'
3) 한선교, 아침 만들기
'여기는 대저내서 마니 오시나요'

Triphone을 사용하여 선택된 문장들을 만들기 위하여 다음과 같은 절차를 거쳤다.

먼저, 수집된 방송 내용 중에서 각 triphone을 포함한 화자의 음성 부분만을 TI(Texas Instruments)사의 TMS 320C30 Evaluation Module(EVM) DSP 보드를 이용하여 약 3초간 16KHz, 16bit 디지탈 신호로 변환하였다. 이 때 동일한 triphone이 여러개일 경우 청음 테스트를 통해 적합한 triphone을 선택하였다. 한편, EVM DSP 보드를 기반으로 운용되는 HSW (Hyper-Signal Workstation) 음성 신호 분석 툴을 적용하여, 샘플링된 음성신호로부터 원하는 triphone을 추출하였다. 추출된 triphone들은 독립된 파일로 저장된 후, 다시 HSW 툴을 통해 각 triphone을 시간축 상에서 변형 없이 연결하는 방법으로 합성 문장을 만들었다. 이상의 triphone 접속에 의해 만들어진 음성파형들이 그림 2, 3, 4에 나타나 있다.

최종 화자를 결정하기 위하여 위에서 만든 3개의 합성 문장들을 들으면서 비교 검토하였다. 우선 전체적으로, 각 합성 문장들에는 각 화자들의 고유한 특성 (예를 들어 음색)이 그대로 살아 있었다. 이는 합성 단위가 triphone임에 기인하는 것으로 사료된다. 이는 또한, 본 연구를 수행하고 있는 가장 근본적인 이유이기도 하다.

6시 내고향의 프로그램으로부터 만든 '잔머리가 마니 도라서 어따주리'는 남성 특유의 굵은 목소리가 살아 있

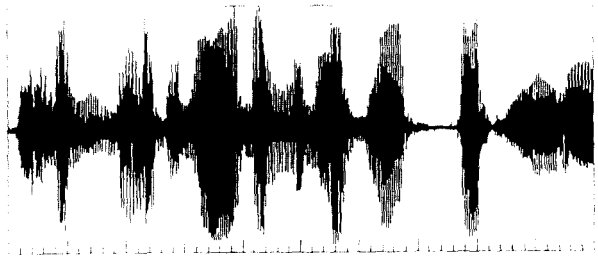


그림 2. 박용호, 6시내고향 "잔머리가 마니 도라서 어따주리"

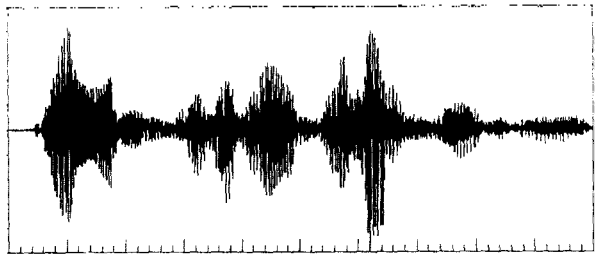


그림 3. 한선교, 아침만들기 "여기는 대저내서 마니 오시나요"

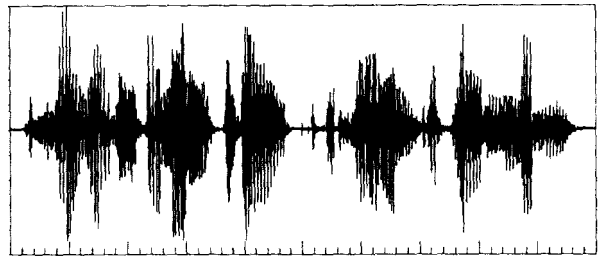


그림 4. 유열, 유열의 음악앨범 "다가서는 절 그대는 보고 인나요"

었으나, 각 음들 사이의 높낮이 변화 폭이 매우 컸다. 물론, 음의 고저 (예를 들어, 피치)를 신호 처리 과정에서 어느 정도 조절할 수는 있으나, 이럴 경우 음색에 적지 않은 영향을 미칠 가능성이 있다. 아침 만들기 프로그램으로부터 얻은 '여기는 대저내서 마니 오시나요'의 합성 문장과, 유열의 음악앨범으로부터 만든 '다가서는 절 그대는 보고 인나요'는, 6시 내고향의 '잔머리가 마니 도라서 어따주리'에 비하여 각 음들 사이의 높낮이 변화 폭이 상대적으로 작았다. 한편 '여기는 대저내서 마니 오시나요'와 '다가서는 절 그대는 보고 인나요'의 경우, 후자의 경우가 전자에 비하여 전체적으로 '안정된' 느낌과 '부드러운' 느낌을 얻을 수 있었다. 이는, 두 프로그램의 성질상, 전자는 talk-show 그 자체의 성격을 띄는 반면에, 후자의 경우는 화자가 청취자들을 대상으로 낭독자의 역할을 했기 때문인 것으로 해석된다. 위에 열거한 분석 결과에 의하여, 유열화자를 본 연구의 triphone 데이타베이스 구축의 최종 화자로 결정하였다.

2.3 최종 화자의 음성 데이타 수집

위에서 설명한 화자 결정 과정에서는 '유열의 음악앨범' 방송을 라디오 공중 파를 이용하여 카세트 테잎에 녹음하였다. 하지만 해당 프로그램이 녹음된 매체를 구할 수 없었고 직접 녹음한다는 것이 불가능하다고 판단하여 최종 화자를 MBC TV '아침 만들기'의 H 화자로 변경하였다. 선정된 최종 화자의 음성 데이타는 MBC TV 방송국으로부터 10시간 분량의 방송 내용이 녹음된 10개의 비디오 테잎을 구입함으로써 비교적 손쉽게 이루어졌다.

수집된 10개 테잎들의 방송 일자는 다음과 같다.

8/2, 8/3, 8/17, 9/25, 10/2, 10/3, 10/4, 10/5, 10/13, 10/18

입수한 10개의 비디오 테잎들 중에서, 실제로 triphone 데이타베이스의 구축을 위하여 6개의 테잎 (즉, 6시간 분량)을 사용하였다. 나머지 4개는 실제 triphone 데이타베이스 구축에는 사용되지 않은 예비용으로, triphone 데이타베이스 구축에 사용된 6개의 테잎과 함께 본 연구의 후반부에 수행된 통계 처리에 사용되었다. Triphone 추출에 사용된 6개의 테잎의 방송 일자는 다음과 같다.

8/2, 8/17, 9/25, 10/2, 10/3, 10/4,

따라서, 예비용으로 비축되어 통계 처리에 사용된 4개의 테잎의 방송 일자는 다음과 같다.

8/3, 10/5, 10/13, 10/18.

참고로, '아침만들기' 프로그램은 2명의 아나운서가 함께 진행하는 아침 talk-show이며, 전반부와 후반부에 걸쳐 1명에서 많게는 5명까지의 초대 손님들과 자연스럽게 대화를 나누는 형식을 취하고 있다. 방송 내용은 주로 연예, 사회문제 등의 흥미 위주의 대화나 시사적인 대화가 가능한 초대 손님과, 생활 상식, 건강 상식 등의 정보를 제공하기 위한 대화가 가능한 초대 손님으로 나누어 진행된다. 또한 방송의 성격상 화자로 선정된 남성 진행자의 음성 이외에 여성 진행자와 초대 손님들의 음성이 섞이는 경우가 많으며, 대화체의 특징인 간투사, 반복 발화, 교정 발화 등의 현상이 자주 나타난다.

Ⅲ. Transcription 작성

Transcription 작성은 대화체 데이타베이스 구축의 전체 과정을 이끌어 가는 가장 중요한 기초 자료가 된다는 점에서 신중을 기해야 하며, 작업 전에 체계적이고 일관성 있는 계획이 꼭 필요한 과정이다. 본 연구에서 작성된 transcription은 각각의 목적에 따라 받아쓰기, 변환용, 발음, 음성기호 그리고 레이블링등 5가지의 transcription으로 구분하여 순서적으로 작성하였다. 각 transcription의(이하 TS) 목적과 내용은 다음과 같다.

3.1 받아쓰기 TS

방송 비디오 테잎을 직접 시청/청취하면서, 화자가 말

한 내용을 맞춤법에 따라 받아 적은 transcription이다. 이를 특별히 구분한 이유는 전체적인 방송 진행 상황에 대한 정보를 기록하고자 하는 것이며, 특별히 음성음향학 또는 음성신호 처리에 대한 지식이 없이도 작성할 수 있는 transcription이기 때문이다. 방송 진행 상황에 대한 정보의 예로는 간단한 방송 상황의 묘사 또는 웃음소리 그리고 동반 진행자의 음성 발화에 대한 정보 등이다.[2] 9월 25일자 방송의 받아쓰기 TS의 예는 다음과 같다.

안녕하십니까? 한선굡니다.(허수경씨 인사)
네 다 알고 계시죠 저희가 누군지는 헌데 정한용씨 피아니스트들도 그, 그렇게 잘 못 칠 때가 있나요.(웃음 ,정한용말)
그러게.
아뇨 좋아요, 그래 보통 한 어 연주회 할 때 몇 번이나 틀리게 되나요.(웃음소리)
아 아 틀리는 것두요(허수경) 뭐요. 윤달 그 신경 굉장히 쓰시데요.
윤달 때문에요, 오늘아 저 팔 월 일일이고, 어 윤달 피하느라고 뭐 이사들도 먼저 가시지 않으면 한달 뒤로 미루시지 않으시며는 아기 낳는 것도 그렇고 (허수경) 그러게 말이에요. 그리고 또 이번 달엔 또 별로 없을 거예요. 그러고 보면.

하지만 대화체의 속성인 빠른 발성 속도 등으로 인해 교육받은 연구원의 경우에도 정확한 발성 문장을 받아적지 못하는 실수를 하였다. 이는 청각적인 한계로 생각할 수 있으며 반복되는 transcription 작성 과정에서 수정되었다. 하지만 미리 방송 대본을 입수하여 작업을 진행한다면 이러한 오류를 최소화할 수 있으며 시간을 절약할 수 있을 것으로 사료된다.

3.2 변환용 TS

변환용 TS는 방송 내용 중에서 최종 화자의 음성만을 선택하여 A/D 변환과정에 사용되는 transcription이다. 따라서 받아쓰기 TS의 정보와 재청취를 통해 잡음, 배경음 또는 다른 화자의 발성 부분이 섞인 부분 등을 제외한 triphone 추출 가능한 부분만을 다시 정리하고 방송 시작 시점을 기준으로 시간을 표시해 주었다. 변환용 TS의 예는 다음과 같다.

(0:39) 네 다 알고 계시죠 저희가 누군 지는 헌데 정한용씨 피아니스트들도 그, 그렇게 잘 못 칠 때가 있나요.
(0:51) 아뇨 좋아요, 그래 보통 한 어 연주회 할 때 몇 번이나 틀리게 되나요.
(1:10) 아 아 틀리는 것두요.
(1:13) 윤달~ 그 신경 굉장히 쓰시데요. 윤달 때문에 오늘이 저 팔 월 일일이고 어 윤달 피하느라고 뭐 이사들도 먼저 가시지 않으면 한달 뒤로 미루시지 않으시며는 아기 낳는 것도 그렇고
(1:27) 그러게 말이에요. 그리고 또 이번 달엔 또 별로 없을 거예요. 그러고 보면.

3.3 발음 TS

발음 TS는, 변환용 TS에 기록된 음성 부분을 발음나는대로 옮겨 쓴 transcription이다. 이는 실제 과형 또는 스펙트로그램을 보면서 적어 나가는 발음과는 정확하게 일치하지 않는다. 그러나, 잘못된 오류는 triphone 추출 과정에서 수정될 것이므로 일단 예상되는 발음대로 옮겨 쓴 transcription이다. 발음 TS는 두 가지 목적을 갖는다. 첫째는, 대화체에서 화자의 발성 또는 발음 습관 등을 파악할 수 있으며, 이는 방송에서 사용되는 용어에 대한 기초 자료로서 사용된다. 둘째는, 자동 변환 프로그램을 사용하여, 음성기호층 레이블링을 위하여 기준형으로 사용되는 음성기호 TS를 얻는데 사용된다.

특히 두 번째 목적을 위해서 발음 TS는 자동 변환이 가능하도록 일정한 포맷 (또는 규칙)을 유지해야 한다. 가장 기본적으로 지켜져야 할 규칙은, A/D 변환 작업이 변환용 TS의 한 문장에 대응하여 한 개의 음성 파일로 저장되었으므로, 음소 표기도 역시 파일 단위로 이루어지도록 각 문장과 문장은 구분되어야 한다는 점이다. 발음 TS의 예는 다음과 같다.

1 (다 알고 계시죠 저희가 누군 지는) 근데 # 정하뇽씨 # 피아니스트들도 # 그, 그러케 잘 못 칠 때가 인나요 # (허수경 웃음)
2 # 아뇨 조아요?(섞임) # 그래 보통 한 #(숨소리) 어 /# 연주회 할 때 며 뻔이나 틀리게 되나요 #
3 아 틀리는 거뚜 (전체섞임)

3.4 음성기호 TS

음성기호 TS는 선택된 화자의 발성 부분에 대한 음성 기호층 표기이며 발음 TS에 자동 변환 프로그램(pro2pho)을 적용하여 생성한다. 각 음성기호의 발생 위치에 대한 정보는 추후 레이블링 TS에서 추가된다. 자동 변환 프로그램 과정에서 사용한 음소 테이블은 표 1에 보인 한국전자통신 연구소에서 제공한 자료를 기초로 하여 작성하였다. 사용한 음소 테이블은 동일한 음소에 대한 변이음, 즉 다른 형태로 발음되는 유성음, 무성음, 마찰음 등의 세분화된 특성으로는 구분하지 않았다[7, 8]. 이는 구축하고자 하는 데이타베이스가 합성을 고려한 것이고, 데이타베이스의 단위가 triphone이므로 각 음소의 변이음을 구체적으로 구분하지 않아도 서로 다른 조합의 triphone이 이미 변이음을 내포한다고 판단되었기 때문이

2. 방송 상황에 대한 정보는 괄호 안에 표시되었다. 간혹 감탄사, 간투사 등의 불명확한 발음이 있음을 나타내기도 한다.

표 1. 음성 기호 표 (계속)

초성	기호
ㄱ	g
ㄴ	n
ㄴㄴ	n2
ㄷ	d
ㄹ	r
ㄹㄹ	l
ㅁ	m
ㅂ	b
ㅅ	s
ㅅㅅ	sh
ㅈ	j
ㅊ	ch
ㅋ	k
ㅌ	t
ㅍ	p
ㅎ	h
ㄲ	g2
ㄸ	d2
ㅃ	b2
ㅆ	s2
ㅉ	j2

중성	기호
아	a
어	v
오	o
우	u
으	eu
이	i
애	ae
에	e
야	ya
여	yv
요	yo
유	yu
얘	ye
와	wa
워	wv
왜, 웨	we
위	wi
왜	wae
의	eui

종성	기호
ㄱ	gs
ㄴ	ns
ㄷ	ds
ㄹ	ls
ㅁ	ms
ㅂ	bs
ㅇ	ng

음소	기호
묵음	#

*예외적인 발음 예('신동호가 여는 아침'으로 작업한 것을 예로…,)
ㄴ2 : '삼백육십'의 발음에서, /뱅뉵/이 될때의 /ㄴ/발음
ㄹ2 : '일루일 내내'의 발음에서, /일래내/가 될때의 초성 /ㄹ/
ㅅ2 : '시간'의 발음에서 /시/가 될때의 구개음화된 /ㅅ/
**ETRI에서 사용한 표기법을 기본으로 하여 작성

다. 하지만 그와 같은 변이음을 포함하는 개념으로 설명할 수 없는 경우가 있다. 대표적인 예로 음소 표기에 추가된 /sh/로서 /ㅅ/ 다음의 /ㅣ/ 모음이 무성음화 되면서 단일 음소로 발음되는 '시간'에서의 /시/(/sh/) 발음을 들 수 있다. 실제로 /ㅅ/ 다음의 짧은 모음들은 빠른 발화 속도에서는 무성음화 되거나 아예 탈락되는 현상이 triphone 추출 작업 동안 자주 관찰되었다[8]. 따라서, ?-/s/-/i/, ?-/s/-/u/ 등의 조합으로는 변이음으로서 무성음화된 모음을 설명할 수 없으므로 음소 표기에 추가하였다. 다음은 발음 TS에서 예를 든 문장을 음성 표기로 변환한 것이다.

SAMPLE FILE-1. TIM
g eu ns d e # j v ng h a n yo ng s2 r # p i a n i s eu l
eu d eu ls d o # g eu g eu r v k e j a ls m o ds ch i ls
d2 ae g a i ns n a yo #
SAMPLE_FILE-2. TIM
a n yo j o a yo # g eu r ae b o t o ng h a ns # v
yv ns j u h we h a ls d2 ae m yv b2 v ns i n a t eu
ls r i g e d we n a yo #
SAMPLE_FILE-3. TIM
a t eu ls r i n eu ns g v d2 u

자동 변환 과정은 한글 코드만을 인식하여 해당 음소의 표기 기호로 변화시킨다. '*', '(', ')', '/', '#' (묵음도 하나의 음성기호로 처리되었다.) 등의 제어 문자로 쓰이는 것을 제외한 나머지 ASCII 코드는 모두 무시된다. 또한 발음 TS에서 '\n' 문자가 나오기 전까지의 문장은 모두 한 개의 샘플링 파일로 간주하여 'SAMPLE_FILE-##.TIM'의 형식으로 파일 명을 붙여서, '방송 일자-##.TIM'으로 명명된 실제 파일과 대응하여 작업할 수 있도록 처리하였다.

이러한 음소 표기의 정의는 triphone 추출 작업에 큰 영향을 미치므로 미리 목적에 맞도록 구체적으로 정의되어야 하지만, 본 연구에서와 같은 대화체에서는 동일한 음소의 수많은 변이음을 발견할 수 있으므로 매우 힘든 작업이다. 물론 사전시험 등의 작업을 통해 검토되었지만 대량의 음성을 다루는 경우에는 모든 경우를 고려한다는 것은 거의 불가능하다고 할 수 있다. 따라서 앞으로 대용량 대화체 음성 데이타베이스의 구축을 위해서는 각 목적에 맞는 적절한 음성기호 표기의 정립이 중요하다고 사료된다.

3.5 레이블링 TS

레이블링 TS는 이후에 수행되는 triphone 추출 작업을 통해서 만들어지는 transcription이다. 또한 최종 결과물인 레이블링 파일을 생성해 내기 위한 transcription 이기도 하다. 레이블링 TS에는, 음성기호 TS에 잘못 기입된 음소를 실제 발음된 음소로 수정하고 탈락된 음소를 삭제한 내용이 포함되어 있다. 또한 발아쓰기, 변환용 TS의 작성 시에 오류로 인해 기입되지 않은 단어 또는 음소를 추가하였다. 물론 각 triphone을 추출하는 동시에 그 시작점과 끝점을 샘플 단위, 또는 시간 단위로 기록하였다.

최종적으로, 레이블링 TS에는 자동 변환 처리를 위한 제어 문자가 추가된다. 이는, 각 샘플 파일 별로 각각의 triphone 위치를 기록한 레이블링 파일을 만들기 위함이다. 레이블링 TS의 예는 다음과 같다.[3]

3. 괄호로 싸여진 부분의 음성은 triphone을 추출하기에 부적합한 음소들이다.

FILE : 0925-1. RAW
{# g eu ns d e # j v ng a yo ng s2 i # p i a n i s t eu d eu ls d u/(# g eu g eu r v k e) j a ls m o ds ch ls d2 ae g a i (ns a yo #)}
{1838 2065 2129 2454 2636 2755 2898 3075 3278 3395 3514 3730 3844 3952 4046/4765 4879 4967 5076 5229 5293 5362}

FILE : 0925-2. RAW
{(# a n yo) j o a yo # g eu (r ae b o t o ng h a ns # v) v yv ns j u we h a ls d2 ae e # m yv bs b2 v ns i n a t eu ls r i (g e d we n a yo #)}
{330 433 640 781/3400 3652 3770 3918 3967 4154 4332 4410 4578 4662 4691 4864 4948 5066}

3.6 방송 매체 및 대화체 특성

이미 지적했듯이 transcription 작성은 대화체 데이타베이스 구축에 있어서 전체 과정의 윤곽과 일정을 결정하는 중요한 기초 자료가 된다. 따라서 각 transcription의 일관성 있는 유지 및 관리는 데이타베이스 구축의 완성도에 중요한 위치를 차지한다고 할 수 있다.

본 연구에서 작성한 변환용 TS, 발음 TS, 레이블링 TS는 각각 일반적인 1, 2, 3차 transcription에 해당하며, 최종 결과물은 텍스트 파일 형태로 저장되는 레이블링 파일이다. 다음은 생성된 레이블링 파일의 일부이다.

FILE : 0925-1. RAW

# g eu : 29838 33523	i s t : 57046 60552
eu ns d : 33523 34562	t eu d : 60552 62403
d e # : 34562 39838	d eu ls : 62403 64156
# j v : 39838 42792	ls d u : 64156 65682
v ng a : 42792 44724	j a ls : 77355 79205
a yo ng : 44724 47046	ls m o : 79205 80634
ng s2 i : 47046 49919	o ds ch : 80634 82403
i # p : 49919 53215	ch ls d2 : 82403 84887
p i a : 53215 55114	d2 ae g : 84887 85926
a n i : 55114 57046	g a i : 85926 87046

지금까지의 transcription 작성 과정에서 보듯이, 본 과제에서 수행하는 데이타베이스 구축은, 선택된 화자가 미리 정해진 대본 또는 문장을 발성하여 녹음하는 경우와는 많은 차이점이 있음을 알 수 있었다. 이는 자연스런 대화체 발성 조건을 획득하는 대신 감수해야 할 여러 어려움이라고 할 수 있다. 지금까지의 transcription 작성 과정을 통해 대화체라는 특성으로 관찰된 현상은 다음과 같다.

1. 빠른 발화 속도로 인해 각 음소의 지속 시간이 매우 짧다.
2. 말 그대로 대화체의 속성으로 다른 화자와의 음성이 섞여 나타나는 경우가 많다.
3. 대부분의 말끝이(~습니다, ~입니까?) 약하면서 빨리 발음된다.
4. 머릿속에서 정리된 또는 대본에 의한 말을 하지 않고, 즉흥적인 애드립일 경우 음의 고저, 음의 톤 등의 변화가 매우 심하고 발음이 불명확해진다.
5. 화자가 항상 일정한 감정 상태를 유지하면서 발음하지 않으므로 전체적인 음색이 일정하지 않다.

위와 같은 어려움은 실제 triphone 추출 과정에서 더욱 어려운 문제를 일으키게 된다. 또한 방송 매체에서 수집한 음성 데이타라는 특성에서 볼 수 있었던 현상은 다음과 같다.

1. 방송 시스템에 의한 shot noise가 발생한다.
2. 초대된 화자가 많을 경우 많은 마이크가 동시에 열리게 된다. 따라서 비례하여 많은 잡음이 섞이게 된다.
3. 비디오 테잎이라는 매체에 의한 noise가 발생된다. 실제로 수집된 10개의 테잎중 1개의 테잎은 다른 테잎과 비교하여 잡음이 특히 심하다. 본 연구에서는 그 테잎은 triphone 추출에서 제외시키고 통계 처리에만 사용하였다.

결과적으로 가장 이상적인 대화체 음성 수집의 조건은 표준어와 명확한 발음을 구사하는 훈련받지 않은 화자가, 저 잡음 환경에서 자연스럽고 적당한 발화 속도로 발성하는 것을 녹음한 후, 화자와 함께 녹음된 음성을 들으면서 transcription을 작성해 나가는 것이라 할 수 있다. 즉 지금까지의 작업에서는 표준어와 명확한 발음을 구사하는 훈련받지 않은 화자가 자연스럽게 발성하는 조건은 만족했지만, 환경은 고르지 못한 저 잡음 환경이었고, 때론 빠르고 정확하지 못한 발화로 인해 transcription 작성 및 이후의 triphone 추출 작업에서 어려움을 겪어야만 했다.

Ⅳ. Triphone 추출

4.1 A/D 변환

이미 작성된 변환용 TS에 의해서, 방송 내용중 triphone이 추출될 부분만을 선택적으로 A/D 변환하는 과정을 거친다. 일단 수집된 비디오 테잎의 음성 데이타들은 DAT로 옮겨지고 다음과 같은 환경과 조건으로 A/D 변환하였다.

DAT Player : Sony, TCD-D3
A/D converter : TI TMS320C30 EVM DSP 보드
SR/해상도 : 16234.677Hz/16bit

이때 주의할 것은 구축될 데이타베이스의 음량을 일정하게 유지해야 한다는 점이다. 즉 일정한 입력 레벨을 유지하면서 A/D 변환을 수행해야만 추후 안정된 데이타베이스가 구축된다는 점이다. 일정하지 않거나, 너무 작거

나 또는 너무 큰 음량으로 녹음된 데이타베이스는 추후 사용에 있어서 불필요한 과정을 거치기 때문이다. 본 연구에서는 약 5000~7000 사이의 amplitude를 갖도록 A/D 변환하였다.

각 샘플링 파일의 이름을 정하는 것 또한 미리 생각해야 할 사항이다. 본 연구에서는 6개의 방송 일자에 대해서 각각 54, 34, 69, 70, 50, 80개의 샘플링 파일을 만들었다. 만들어진 샘플링 파일들은 triphone을 추출할 수 있다고 판단된 화자의 음성 외에 여러 가지 신호가(다른 화자의 음성, 웃음소리, 잡음) 포함된 파일들이다. 따라서 모든 파일들은 음성 파형 편집 툴이나 스펙트로그램을 볼 수 있는 툴을 사용하여 불필요한 부분은 삭제하고 필요하다면 전처리와 같은 변환 파일을 만드는 등의 많은 처리가 요구된다. 이러한 처리를 위해 각 샘플 파일 명은 각 파일간의 혼동을 막고 다른 방송 일자의 샘플링 파일들과의 구분이 쉬워야 한다.

본 연구 과정에서는 '방송 일자-샘플 번호'의 포맷으로 파일 명을 붙여 주었다. 가령 9/25의 방송 내용을 기록한 변환용 TS에서 5번째로 기록된 문장을 A/D 변환할 때에는 '0925-5'라는 파일 명을 붙여 주었다. 또한 이 파일을 전처리한 파일은 파일 명의 끝에 'p'를 붙임으로써 혼란을 막고 일관성을 유지할 수 있었다.

4.2 Triphone 추출

지금까지의 작업을 통해 얻어진 것은 음성 표기 테이블, 파일 단위의 샘플링 파일, 파일 단위의 샘플링 파일에 대한 음성 표기 transcription이다. 이 세 가지 자료를 이용하여 triphone을 추출하게 된다. Triphone 추출 과정에서는 음성 표기 테이블에 정의된 음소만을 추출하게 된다. 일단 작업이 시작되면 일관성을 유지해야 하므로 새로운 음성기호를 추가하거나 분류할 수 없게 된다. 따라서 사전시험, transcription 작성과 같은 사전 준비 작업을 통해서 충분히 검토되어야 한다.

Triphone은 '중심이 되는 임의의 음소와 앞 음소의 후반부 및 뒤 음소의 전반부를 포함하는 단위'라고 정의할 수 있는 문맥 종속(CD, Context-Dependent) 음소 단위이다. 따라서 같은 /s/(/ㅅ/) 음을 포함한 triphone이라 해도, 전/후에 발생된 음소에 따라서 전혀 다른 triphone으로 분류될 수 있다. 이점은 바로 세분화된 각 음소(음성) 표기 기호가 필요하지 않다는 것을 설명해 준다. 즉 중심이 되는 음소의 여러 가지 변이음을 다른 표기로 구분하지 않더라도, 전/후에 발생된 음소에 따라 이미 중심 음소는 다른 조합의 동일한 중심 음소와는 다른 변이음으로 생각할 수 있다. 이러한 점은 자연성을 추구하는 합성 시스템에서 triphone 사용의 당위성을 간접적으로 설명한다고 할 수 있다.

위와 같은 기준으로, 본 연구에서는 TI TMS320C30 EVM을 기반으로 운용되는 HSW 툴과, IBM-PC의 SoundBlaster를 기반으로 운용되는 윈도용 음성 분석 툴인 Cool을[4] 보조 툴로서 사용하여 triphone을 추출하였다. Triphone에서 앞 음소의 후반부 또는 뒤 음소의 전반부의 위치는 각각 안정된 부분의 중심점으로 결정하였다. 또한 음성기호 TS를 기준으로 3개씩의 음소를 차례대로 하나의 triphone으로 추출하였다. 따라서 각 triphone은 전/후의 triphone과 중복되지 않는다. Triphone을 구성하는 각 음소의 위치는 시간축 파형과 전처리된 파형을 160개 샘플의 프레임 크기, 80개 샘플의 분석 주기로 분석된 스펙트로그램을 함께 보면서 확인하였다. 이때 음성기호 TS의 음성 기호 표기를 확인하는 작업도 병행되었다.

본 연구에서는 전처리한 신호의 스펙트로그램을 사용하였다. 이는 빠른 발화 속도와 전이 속도를 갖는 대화체의 신호에서 각 음소의 경계를 좀 더 쉽게 찾기 위함이다. 만약 triphone을 구성하는 앞 또는 뒤 음소의 정확한 위치를 찾지 못할 경우에는 정확한 위치를 찾을 수 있는 음소부터 다시 하나의 triphone으로 취급하였다.

이와 같은 triphone 추출 작업 동안에, 낭독체와 같은 적당한 발화 속도 환경에서는 볼 수 없는 많은 동시조음에 의한 현상들을 관찰할 수 있었다. 이미 연구된 바 있지만 대표적인 예로 몇 가지를 들어보면, 성문 마찰음 /ㅎ/은 모음 사이에서 약화되거나 대부분 찾기 어려울 정도로 탈락되는 현상, /ㅅ/, /ㅊ/, /ㅋ/ 등의 마찰음 뒤에서 반모음뿐이 아니라 일부 모음과도 융화하여 하나의 음소로 발음되는 현상 등이다[3, 8].

이러한 대화체의 특성을 모두 고려하면서 정확한 음소 위치를 찾아내는 것은 음성음향학적인 전문적인 지식이 바탕이 되어야 한다고 생각된다. 하지만 일관된 작업을 위해서 대화체의 특성을 고려한 다음과 같은 몇 가지 기준을 세울 수 있었다.

1) 앞 음소의 후반부와 뒤 음소의 전반부는 파형에서 그 음소가 안정된 구간을 보일 경우 가운데 지점을 경계로 한다.
2) 파형 상에서 음소의 위치가 확인되지 않을 경우에는 스펙트로그램으로 그 경계를 찾고 역시 중심 위치는 음소의 가운데 지점을 경계로 한다.
3) 묵음도 하나의 음소로 취급하며 원칙적으로 60msec 이상의 시간 동안 에너지가 관찰되지 않을 경우 묵음으로 간주한다.
4) 초성에서 발생하는 /ㅍ/, /ㅌ/ 등의 파열음 앞의 묵음은 파열음의 일부로 생각하며, 역시 종성에 위치하는 /ㄱ/, /ㄷ/ 등의 폐쇄음에 수반되는 묵음 역시 그 음소로 간주한다. 따라서, 그와 같은 음소에서 묵음만을 독립적으

4. Syntrillium Software Corporation제작. 본 연구에서는 DEMO 버전을 사용하였다.

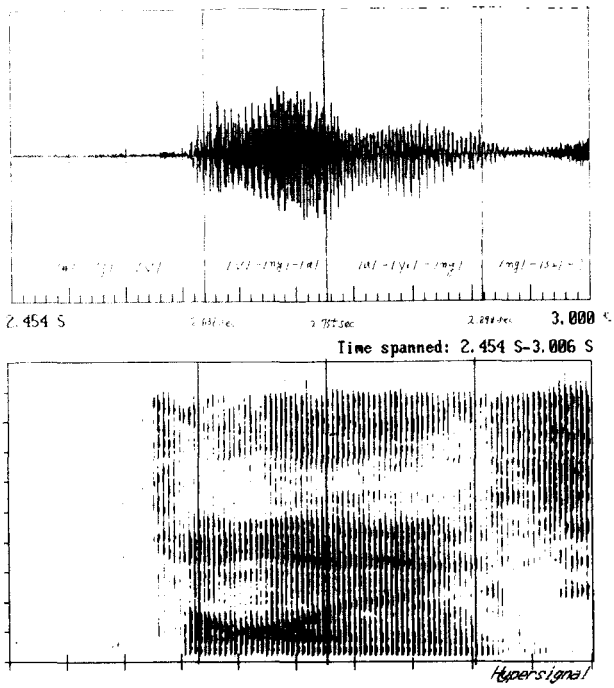


그림 5. "#정한용씨"의 레이블링(위 : 시간축상의 파형, 아래 : Pre-emphasis 된 파형의 스펙트로그램)

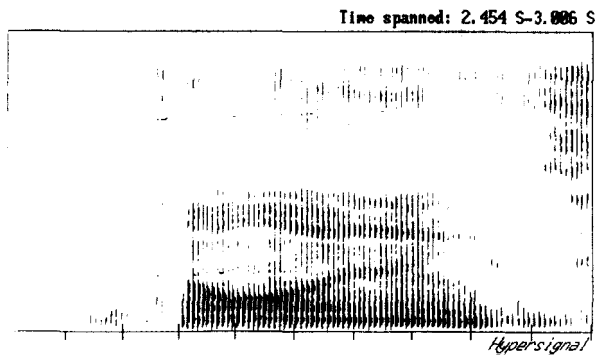


그림 6. Pre-emphasis 되지 않은 파형의(그림5. 위) 스펙트로그램

로 취급하지 않는다.

5) 스펙트로그램에서도 융화되거나 전이되어 관찰하기 어려운 음소의 결합은 굳이 나누지 않고 하나의 음소로 취급하거나 무시하는 것을 원칙으로 한다.

6) 이중모음과 같은 지속적인 전이 현상을 수반하는 경우는 확실한 전이 구간이 관찰될 경우에만 전이의 중간 지점을 음소의 중간 지점으로 간주하고 이외에는 하나의 음소로 취급한다.

이상의 개괄적인 기준을 바탕으로 triphone을 추출한 시간축, 주파수축 상에서의 예를 그림 5에 나타내었다. 그림 5의 위는 "#정한용씨"를 발음한(/#정아용씨/) 파형의 일부를 보여주고 있다. 이 파형은 추출 작업이 비교적 어려운 예로써 시간축 상의 파형만을 보고 각 음소를 구분하기가 매우 어렵다. (특히 /h/, /ns/ 음은 거의 발견되지 않음을 볼 수 있다.) 따라서 그림 5의 아래에서 보인 스펙트로그램을 위주로 추출 작업이 이루어졌다. 참고로 그림 6은 전처리되지 않은 원래 신호의 스펙트로그램을 나타내었다. 그림 5의 아래에 보인 스펙트로그램과 비교하여 변이음을 거의 구분하지 못함을 알 수 있다.

실질적으로 가장 중요한 과정인 triphone추출 작업은 많은 시간, 인력 그리고 음성 및 음향학에 대한 지식이 모두 만족되어야만 완벽하고 정확한 분할이 이루어지며, 이것은 결국 구축된 데이타베이스의 완성도에 큰 영향을 끼친다고 할 수 있다. 하지만, 지금까지의 작업에서 볼 수 있듯이 이에 못지 않게, 일련의 작업들이 상호 유기적인 관계를 가지고 진행되므로, 음소 표기 테이블의 결정, transcription 작성, A/D 변환 그리고 triphone 추출 작업들이 일관된 계획으로 진행되어야 함을 강조하고 싶다.

V. DB 구축과 통계 처리

최종적으로 triphone추출 작업을 통해 완성된 데이타베이스의 전체적인 정보는 다음과 같다.

방송일자 (월/일)	샘플링 파일 개수 (개)	파일 총 크기 (bytes)	triphone 개수 (개)
8/2	39	6,564,520	294
8/17	18	7,119,772	377
9/25	43	11,942,862	501
10/2	47	10,470,110	265
10/3	33	10,472,766	205
10/4	43	9,862,308	159
평균	37	9,405,389	300
합계	223	56,432,338	1,801

첫 번째 통계 처리는, 6시간 분량의 방송에서 추출된 triphone들 중에서 중복된 것을 제외한 triphone의 가짓수를 측정하는 것으로서, 총 1,015가지의 triphone이 추출되었다. 따라서 약 43.6% 의 triphone이 중복되어 추출되었음을 알 수 있다.

두 번째 통계 처리는, 예비 분으로 수집된 4시간 분량에서 발음 및 음성기호 TS를 기준으로 추출 가능한 triphone의 가짓수를 알아보는 것이다. 이론적으로, 모든 가능한 경우의 수를 고려할 때, 약 3,714가지의 중복되지 않는 triphone 추출이 가능하며, 이중 6시간 분량과 중복된 triphone을 제외하면 2,699가지의 새로운 triphone이 포함되어 있음을 알 수 있었다. 하지만, 이러한 결과는 triphone을 추출할 때 음소의 경계를 중복시키지 않는 본 연구 과정에서의 원칙을 적용하였으므로, 실제로는 더욱 많은 triphone이 발생될 수 있다. 물론 이때 실제 뽑아 낼 수 있는 triphone의 수는 첫 번째 통계 결과를 통해 볼 때

30%에 못 미칠 것으로 예상된다. 이는 각종 잡음, 불명확한 발음 등의 이유로 추출 불가능한 음성이 발생되기 때문이다. 따라서, 예비 분의 4시간 분량의 음성 데이타에서 triphone을 뽑는다고 가정한다면, 첫 번째 통계 결과를 적용할 때 실제 추출할 수 있는 trip hone의 가짓수는 약 676개 (2,699개의 약 26%) 정도로 예측된다.

Ⅵ. 결 론

본 연구는 합성 시스템을 위한 triphone 단위의 남성 음성 데이타베이스 구축을 목표로 하였으며, 대화체 음성의 수집을 위하여 방송 매체를 이용하였고 적절한 프로그램 및 화자 선정 과정을 거쳤다. 총 10시간 분량의 방송을 수집하였고, 그 중에서 6시간 분량에 대해서 triphone 단위의 분할 및 음성기호층 레이블링 작업을 수행하였다. 결과적으로 총 1,801개, 가짓수로는 1,015개의 triphone을 추출하였다. 예비 분으로 수집된 4시간 분량에서는 이론적으로 2,699가지의 triphone을 더 추출할 수 있음을 통계 처리를 통해 알아냈다.

본 연구를 통해, 방송 매체를 통한 대화체 음성 데이타베이스를 구축하는데 있어서 다음과 같은 문제점들을 제기할 수 있다.

1) 적절한 대화체를 수집하기 위한 방법이 정립되어야 한다.
2) 방송 매체를 통한 대화체 음성 데이타 수집이 가장 객관적이고 경제적인 방법이 될 수 있는 방안을 찾아야 한다.
3) 대화체의 특성을 충분히 이해하는 과정이 선행되어야 한다.
4) 빠른 발화 속도와 동시 조음 현상을 고려하기 위한 일관된 원칙이 사전에 세워져야 한다.
5) 대화체를 고려한 자동 묵음 검출 기법 및 유/무성음 검출기를 비롯한 자동 레이블러가 개발되어야만 시간과 인력을 절약할 수 있다.
6) 음성 데이타 수집 단계에서 triphone 추출의 효율을 따질 수 있는 기준이 마련되고 측정되어야만, 경제적인 데이타베이스 구축이 가능하다.

이상의 문제점들을 극복하기 위한 향후 연구 방향은 크게 두 가지로 나눌 수 있다.

첫째는 낭독체 음성을 이용하여 triphone 데이타베이스를 구축하는 것이다. 이 연구에서는 잡음, 빠른 발화 속도, 발음의 불명확 및 빠른 운율 변화 등의 대화체 고유의 특성을 고려하지 않음으로써 triphone 데이타베이스의 타당성과 자동/반자동 레이블러 및 합성 시스템 등의 응용 연구를 진행하는 것이 목표이다. 또한 동시에 대화체 음성 데이타베이스 구축에 적용될 기준과 보완 점을 연구할 수 있다고 사료된다.

또 하나의 연구 방향은 궁극적인 음성 인식/합성에 사용될 수 있는 대화체를 이용한 triphone 데이타베이스를 지속적으로 보완해 나가면서 구축하는 것이다. 이 연구는 낭독체를 이용한 triphone 데이타베이스 구축과 병행되어 진행되어야 하며 운율 정보 등을 포함한 다층 레이블링(multiple-labeling)기법이 연구될 수 있을 것이다.

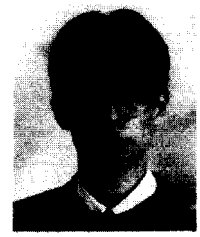
참 고 문 헌

1. 이용주, 이정철, 김경태, "음성 데이타베이스의 구축에 관하여", 한국음향학회지, 제7권, 제5호, pp. 5-13, 1988년
2. 이용주, 임연자, 한남용, 최준혁, 정유현, "ETRI의 음성 및 텍스트 데이타베이스의 구축 현황", 제 1회 ETRI 음성, 언어 및 음향정보처리 워크샵, pp. 161-177, 1993년, 4월.
3. 이호영, "대화체 음성 및 운율 DB", 제12회 음성통신 및 신호처리 워크샵 논문집, pp. 298-301, 1995년, 6월
4. A. Kurenmatsu and K. Takeda, "ATR Japanese speech database as a tool of speech recognition and synthesis", Proc. ESCA '89, pp. 2. 3. 1-4, 1989
5. S. Itahashi, "Recent Speech Database Projects in Japan", Proc. ICSLP 90, pp. 1081-1084, 1990
6. 정재호, "남(여)음성 triphone 데이타베이스 구축에 관한 연구", 한국전자통신연구소 최종 보고서, 인하대학교 산업과학기술 연구소, 1995년, 12월
7. 정국, "음성인식/합성을 위한 기본 개념과 표기법의 정립", 제11회 음성통신 및 신호처리 워크샵, pp. 37-41, 1994년, 10월.
8. 이호영, 지민재, 김영송 "동시조음에 의한 변이음들의 음향적 특성", 한글 제220호 별쇄본, 1993년 6월

▲김 유 진(Yu-Jin Kim) 1969년 11월 22일생

1995년 2월 : 인하대학교 전자공학과 졸업(공학사)
1995년 3월~현재 : 인하대학교 대학원 전자공학과 석사과정

▲백 상 훈(Sang-Hoon Baek) 1971년 10월 24일생



1994년 2월 : 인하대학교 전자공학과 졸업(공학사)
1996년 2월 : 인하대학교 대학원 전자공학과 졸업(석사)
1996년 3월~현재 : 서울이동통신 중앙연구소 연구원

▲한 민 수(Minsoo Hahn) 1956년 11월 23일생

1979년 2월 : 서울대학교 전기공학과 (B.S.)
1981년 2월 : 서울대학교 전기공학과 (M.S.)
1989년 12월 : University of Florida, Eng.(Ph.D.)
1982년 4월~1985년 8월 : 한국표준과학 연구원
1990년 2월~현재 : 한국전자통신연구소 음향통신연구실 실장
※주관심분야 : 음성분석, 합성, 인식 및 음향신호처리

▲정 재 호(Jae Ho Chung)

1982년 : 美國 University of Maryland (학사)
1984년 : 美國 University of Maryland (석사)
1990년 : 美國 Georgia Institute of Technology(박사)
1984년~1985년 : 美國 국방성 산하 해군연구소, 신호처리실, 연구원
1991년~1992년 : 美國 AT&T Bell Laboratories, 음성신호처리 연구실, 연구원
1992년~현재 : 인하대학교 공과대학 전자공학과, 현(부교수)