다이폰 군집화와 개선된 스펙트럼 완만화에 의한 음성합성

장 효 종[†]·김 관 중[††]·김 계 영[†††]·최 형 일[††††]

요 약

본 논문에서는 단위음소들의 연결을 통한 음성합성 방법에 관하여 기술한다. 이때, 발생하는 가장 큰 문제점은 두 단위음소 사이의 연결부분에서 불연속이 발생하는 것이며, 특히 다른 화자로부터 녹음한 단위음소의 연결에서 불연속이 많이 발생한다. 이 문제를 해결하기 위하여 본 논문에서는 군집화된 다이폰을 이용하며, 포만트 궤적과 스펙트럼의 분포특성을 사용할 뿐 아니라 인간의 청각적인 특성을 반영하여 스펙트럼을 완만화하는 방법을 제안한다. 즉, 제안하는 방법은 단위음소 연결구간의 스펙트럼 분포특성의 유사도를 사용하여 단위음소들을 군집화하고 단위음소의 연결 구간에서 인간의 청각신경 특성을 고려하여 완만화의 양과 범위를 결정한 다음, 두 다이폰 경계의 스펙트럼 분포를 시간에 따라 가중치를 다르게 주어 스펙트럼 완만화를 수행한다. 이 방법은 불연속을 제거하며 완만화로 인하여 발생할 수 있는 음성의 왜곡을 최소화한다. 제안하는 방법의 성능을 평가하기 위하여 5명으로부터 녹음한 20개의 문장 중에서 추출한 500여 개의 다이폰을 사용하여 실험을 수행하였다.

Speech Synthesis using Diphone Clustering and Improved Spectral Smoothing

Hyo-Jong Jang[†] · Kwan-Jung Kim[††] · Gye-Young Kim[†††] · Hyung-Il Choi[††††]

ABSTRACT

This paper describes a speech synthesis technique by concatenating unit phoneme. At that time, a major problem is that discontinuity is happened from connection part between unit phonemes, especially from connection part between unit phonemes recorded by different persons. To solve the problem, this paper uses clustered diphone, and proposes a spectral smoothing technique, not only using formant trajectory and distribution characteristic of spectrum but also reflecting human's acoustic characteristic. That is, the proposed technique performs unit phoneme clustering using distribution characteristic of spectrum at connection part between unit phonemes and decides a quantity and a scope for the smoothing by considering human's acoustic characteristic at the connection part of unit phonemes, and then performs the spectral smoothing using weights calculated along a time axes at the border of two diphones. The proposed technique removes the discontinuity and minimizes the distortion which can be occurred by spectrum smoothing. For the purpose of the performance evaluation, we test on five hundred diphones which are extracted from twenty sentences recorded by five persons, and show the experimental results.

키워드 : 다이폰 군집화(Diphone Clustering), 포만트 궤적(Formant Trajectory), 스펙트럼 완만화(Spectral Smoothing), 청각적인 특성(Acoustic Characteristic), 음성합성(Speech Synthesis)

1. 서 론

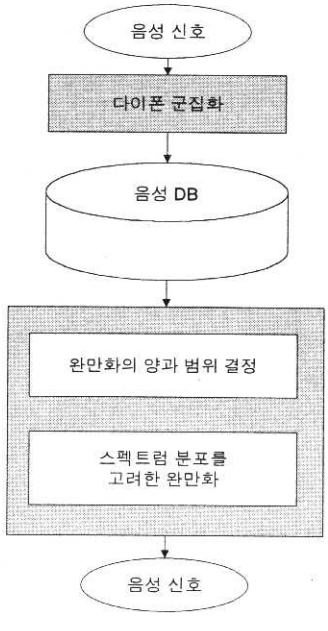
음성합성 시스템에 의하여 생성된 합성음이 자연스럽지 못한 경우는 청자로 하여금 이질감을 느끼게 하고, 부정확한 경우는 정확한 의사전달을 방해하게 된다. 따라서, 음성합성 시스템의 궁극적인 목적은 정확한 정보를 전달하기 위하여 보다 자연스럽고 정확한 합성음을 생성하는데 있다. 현존하는 음성합성 시스템 대부분은 기본적으로 음성 데이터베이스로부터 단위음소를 추출하고 이를 연결하여 합성음을 얻는 방법을 취한다. 이 접근 방법의 대표적인 문제점은 단위음소의 연결 부분에서 불연속이 발생하는 것이다. 그 원인은 데이터베이스에 있는 단위음소들이 모든 경우에 대한 문맥적인 차이나 변화들을 나타낼 수 없기 때문이다. 특히 다른 화자로부터 녹음된 단위음소들을 연결하여 음성을 합성하고자 하는 경우에는 더 많은 불연속이 발생하게 된다.

단위음소의 연결 부분에서 발생하는 불연속을 해결하기 위한 기존의 주요 연구에서는 트라이폰과 같은 큰 합성 단위를 사용하는 방법[1], 스펙트럼 불연속이 최소화된 연결 위치를 찾아 이를 연결하여 불연속을 최소화하는 방법[2], 파형 또는 스펙트럼 완만화 등을 통해 스펙트럼의 불연속을 제거하는 방법[3] 등이 있다. 이러한 방법들의 문제점을 살펴보면 다음과 같다. 첫 번째 방법은 불연속이 완전히 없

※ 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음.
† 준 회 원 : 숭실대학교 대학원 컴퓨터학과
†† 정 회 원 : 한서대학교 컴퓨터정보학과 교수
††† 종신회원 : 숭실대학교 컴퓨터학부 교수
†††† 종신회원 : 숭실대학교 미디어학부 교수
논문접수 : 2003년 5월 2일, 심사완료 : 2003년 8월 8일

어지는 것이 아니라 빈도가 줄어들 뿐이다. 또한 큰 합성 단위를 쓰기 때문에 필요한 자료양이 많아 데이터베이스의 크기를 증가시키는 단점이 있다. 두 번째 방법은 연결 부분의 포만트 궤적이 수평이 아니라는 가정에 근거한다. 이 가정은 스펙트럼의 불연속이 두드러지게 나타나는 모음에서는 포만트의 궤적이 수평으로 나타나기 때문에 연결 위치의 변화만으로는 불연속을 제거할 수 없는 단점이 있다. 세 번째 방법은 완만화를 수행 할 때 불연속 정도에 상관없이 고정적으로 수행하기 때문에 그 결과를 신뢰하기가 어렵다. 이 외에도 기존의 방법 중에는 단위음소의 연결 부분에서 나타나는 스펙트럼의 차이를 수정하여 불연속을 제거하는 여러 시도가 있다[4, 5]. 그 중에 몇몇 방법은 스펙트럼 완만화가 오히려 합성의 질을 떨어뜨리는 결과를 보여주기도 한다. 예를 들면, 연결 부분에서 갑자기 나타나거나 혹은 사라지는 스펙트럼의 피크와 같은 스펙트럼의 왜곡이 생기는 경우이다[4]. 이러한 단점을 해결하기 위해 데이터베이스로부터 이상적인 통합 유닛을 추출하여 이를 스펙트럼 스무딩에 이용하는 방법이 있다[5]. 그러나 이 방법 또한 이상적인 통합 유닛을 추출하는 것이 어려운 문제점이 있다. 본 논문에서는 군집화를 통한 다이폰을 단위 음소를 사용하여 포만트 궤적 뿐 아니라 스펙트럼의 분포특성과 인간의 청각적인 특성을 반영하여 스펙트럼을 완만화하는 방법을 제안한다.

제안하는 방법은 단위음소 연결구간의 스펙트럼 분포특성의 유사도를 사용하여 단위음소들을 군집화하고 단위음소의 연결 구간에서 인간의 청각신경 특성을 고려하여 완만화의 양과 범위를 결정한 다음, 두 다이폰 경계의 스펙트럼 분포를 시간에 따라 가중치를 다르게 주어 스펙트럼 완만화를 수행한다. 이 방법은 불연속을 효과적으로 제거하며 완만화로 인하여 발생할 수 있는 음성의 왜곡을 최소화한다. (그림 1)은 본 논문에서 제안하는 음성합성 시스템의 개요도이다.



(그림 1) 음성합성 시스템 개요도

본 논문의 구성은 다음과 같다. 제 2장에서는 단위음소 연결구간의 스펙트럼 분포특성의 유사도를 사용하여 단위음소들을 군집화하는 과정에 대해서 서술하고 제 3장에서는 인간의 청각신경특성을 반영한 완만화의 양과 범위 결정에 대해서 기술하며 제 4장에서는 결정된 완만화의 양과 범위를 사용하여 포만트 궤적과 포만트 주변의 스펙트럼 분포특성을 반영하여 완만화를 수행하는 방법에 관하여 설명한다. 제 5장에서는 실험결과를 보이며, 마지막으로 제 6장에서는 결론 및 향후연구에 관하여 논술한다.

2. 다이폰 군집화

이 절에서는 본 논문에서 제안하고 있는 개선된 스펙트럼 완만화의 전처리 단계로써 기존의 다이폰 군집화 방법을 사용한 배경과 그 내용에 대하여 기술한다.

이 방법의 경우 음성을 합성할 때 두드러진 스펙트럼의 불연속이 나타나는 모음(/a/, /i/, /u/)에 초점을 맞추어 문제를 해결하고 있다[6, 7]. 먼저 다이폰 연결 부분에서의 스펙트럼이 유사한 다이폰들을 군집화 하여 적절한 크기의 데이터베이스를 구축한다. 이때 각 군집을 대표하는 자음을 중심으로 오른쪽 다이폰 군집들을 재구성 함으로써 데이터베이스의 크기를 줄일 수 있었다. 또, 서로 다른 화자로부터 추출된 단위음소 연결구간의 스펙트럼 분포특성의 유사도를 기준으로 군집화를 수행한 단위음소를 합성에 사용함으로써 스펙트럼 완만화에서 발생할 수 있는 왜곡을 최소화할 수 있었다. 따라서 본 논문에서는 완만화의 전처리 단계로 다이폰 군집화를 수행하고 여기에 개선된 완만화 방법을 적용하여 보다 자연스러운 음성합성을 수행하고자 한다.

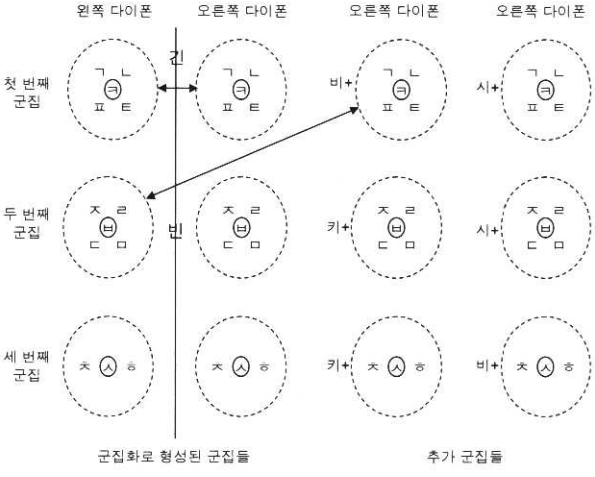
군집화 단위로는 문맥에 민감한 다이폰을 사용한다. 군집화는 LBG 알고리즘을 사용하며 다이폰 경계의 유사도를 기준으로 수행한다. 유사도는 Kullback-Leibler 거리 공식에 의해 계산되며 수식은 다음과 같다[6].

$$KL(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx \qquad (1)$$

식 (1)에서 f(x)와 g(x)는 각각 두 단위음소의 경계에 있는 스펙트럼의 주파수 강도의 확률분포함수이다. 이 두 함수의 관계 f(x) = g(x) 이면 두 경계의 유사도가 매우 높음을 의미한다. 즉, 실제적인 KL-거리는 연결 부분에서 주파수 강도의 밴드별 평균과 표준편차 사이의 차이를 누적함에 의하여 산출된다. 다음에 나오는 (그림 2)는 군집화의 예를 보여준다.

(그림 2)에서 군집화로 생성된 군집들과 추가된 군집들이

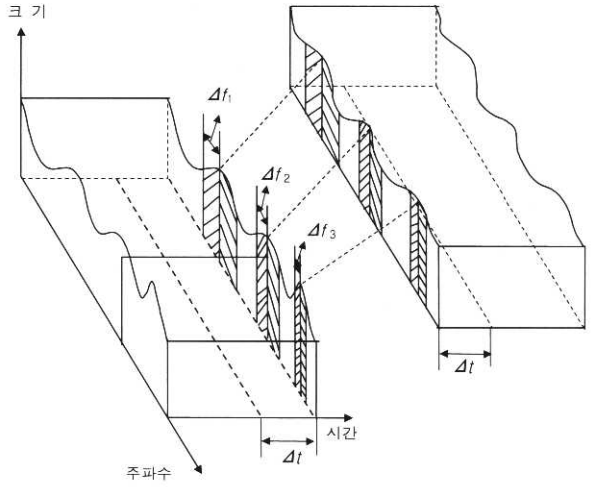
있는데, 군집화로 생성된 군집은 모음 'ㅣ'에 대하여 군집화한 결과로 생성된 3개의 군집을 나타낸다. 예를 들어 '긴'이라는 글자를 합성으로 생성할 때, 모음 'ㅣ'에 대하여 생성된 첫 번째 군집에서 'ㄱ'과 'ㄴ'을 취하면 불연속이 가장 적게 일어나도록 연결이 가능하다. 하지만 두 번째 예인 '빈'의 경우 모음 'ㅣ' 앞의 'ㅂ'과 'ㄴ'은 같은 군집에서 선택할 수 없다. 이를 해결하기 위해서 각 군집의 대표 자음을 기준으로 그 군집에 없는 자음들을 군집화하여 해당 군집에 추가해서 사용하게 된다.



(그림 2) 모음 'ㅣ'에 대한 다이폰 군집화의 예

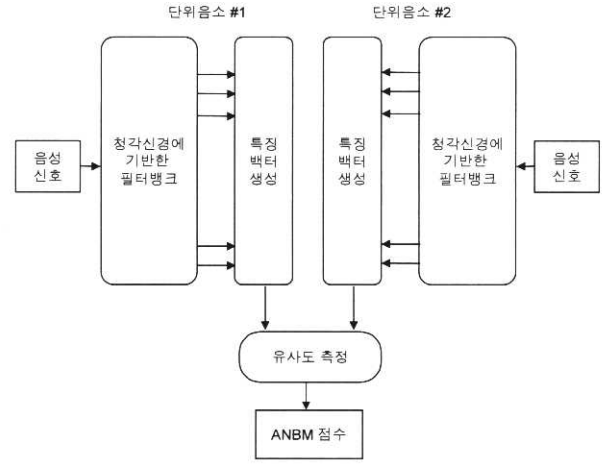
3. 인간의 청각신경 특성을 반영한 완만화의 구간길이와 주파수 범위 결정

이 절에서는 인간의 청각신경 특성을 반영하여 스펙트럼 완만화를 위한 구간길이와 주파수 범위를 불연속의 정도에 적응적으로 결정하는 방법에 대하여 기술한다. (그림 3)에서는 구간질이 Δt와 주파수 범위 Δf의 예를 보인다.



(그림 3) 완만화를 위한 구간길이와 주파수 범위

인간의 귀에 있는 달팽이관의 해부학적 구조에 의하여 인간은 고주파의 변화보다는 저주파의 변화에 더 민감하다는 사실을 알 수가 있다[8]. 따라서 이러한 특성을 스펙트럼 완점수[9]라는 척도를 이용하고자 한다. 즉, ANBM 점수는 인간의 귀 내에 있는 청각 신경세포(Basilar Membrane)의 분포가 모든 대역에 대하여 일정하지 않은 것을 반영하여 단위음소들의 연결부분에서 발생하는 불연속의 정도를 나타내는 척도이다 (그림 4)는 ANBM 점수를 산출하는 과정을 나타낸 것이다.



(그림 4) ANBM 점수를 산출하는 과정

ANBM 점수를 산출하는 방법은 다음과 같다. 먼저, 두 개의 음성신호 각각을 청각신경에 기반한 필터뱅크를 사용하여 대역을 나눈 다음, 각 대역별로 최대강도를 가지는 두 주파수의 차를 합하여 얻어진다. 이 과정을 수식으로 표현하면 다음의 식과 같다.

$$d(x, y) = \sum_{k=1}^{N} |x_k - y_k| \qquad (2)$$

여기서, x와 y는 두 개의 단위음소에 대한 음성신호이고, x_k와 y_k는 k 대역에서 최대강도를 가지는 두 개의 주파수이다. 식 (2)의 의미는 주어진 음성신호에서 모든 채널에 대한 최대 크기를 가지는 주파수의 차를 모두 누적시킨 것이므로, ANBM 점수가 낮을 경우는 불연속에 대한 청각인지도가 낮은 상태, 높은 경우는 불연속에 대한 청각인지도가 높은 상태를 의미한다. 필터뱅크는 참고문헌[8]에서 제시된 방법을 사용한다. 이 방법에 의하여 나누어진 주파수 대역은 <표 1>과 같다.

본 논문에서는 위에서 구한 ANBM 점수를 스펙트럼 완만화의 구간길이와 주파수 범위을 결정하기 위하여 아래와 같이 3단계로 분류한다.

- ANBM 점수 High : ANBM값 = 4, 완만화 수행 구간

30ms

- ANBM 점수 Middle : ANBM값 = 2, 완만화 수행 구간 20ms
- ANBM 점수 Low : ANBM값 = 1, 완만화 수행 구간 10ms

〈표 1〉 청각신경에 기반한 필터뱅크

대역번호	중심주파수(Hz)	대역폭(Hz)
1	50	0~100
2	150	100~200
3	250	200~300
4	350	300~400
5	450	400~510
6	570	510~630
7	700	630~770
8	840	770~920
9	1000	920~1080
10	1170	1080~1270
11	1370	1270~1480
12	1600	1480~1720
13	1850	1720~2000
14	1250	2000~2320
15	2500	2320~2700
16	2900	2700~3150
17	3400	3150~3700

상기와 같이 등급화된 ANBM 값을 사용하여 완만화가 수행되는 주파수의 범위는 식 (3)을 통하여 산출된다. 여기서, W는 포만트 f_0가 존재하는 뱅크의 대역폭이다. 즉, Δf 값은 포만트 f_0가 존재하는 뱅크의 대역폭과 등급화된 ANBM 값에 따라 적응적으로 결정된다. 또한 실험에서 ANBM을 사용하지 않은 경우에 대해서 식 (2), 식 (3)이 계산될 때 ANBM값은 1이 사용되었다.

$$\Delta f = W/ANBM\text{값} \tag{3}$$

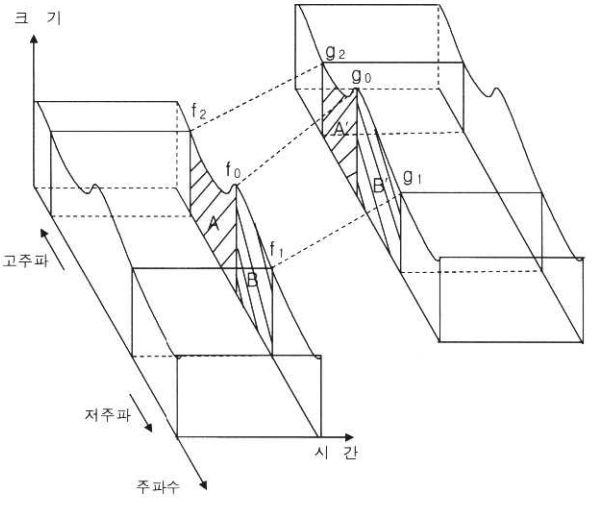
$$\int_{f_0}^{f_0+\Delta f} f(x)dx = C \tag{4}$$

식 (3)에서 결정된 Δf는 식 (4)에서 스펙트럼 분포를 이용한 완만화에 사용되는 단위 넓이 C값을 결정하는데 사용된다.

3. 비선형적 가중치에 의한 스펙트럼 스무딩의 완만화

이 절에서는 2장에서 계산된 ANBM 점수를 이용하여 적응적으로 결정된 대역에 대하여 포만트 궤적과 그 주위의 분포를 고려하여 스펙트럼 완만화를 수행하는 방법에 대하여 기술한다. 스펙트럼 완만화는 보간 포인트를 결정하는 단계와 보간값을 산출하는 단계로 구성된다. 먼저, 포만트를 추출하는 방법에 관하여 설명한 다음, 보간 포인트를 결정하는 방법과 보간값을 산출하는 방법 순으로 기술한다.

포만트 주파수는 특정한 음성이 발성될 때 성도의 공명주파수를 의미하는데, 이를 단순히 포만트라고도 한다. 이 공명주파수를 스펙트럼 관점에서 보면 봉우리 형태로 나타나게 되고, 저주파에서부터 나타나는 순서대로 제 1포만트, 제 2포만트 등으로 표기한다. 포만트의 중심주파수를 산출하는 방법에는 퓨리에 변환이나 필터뱅크의 출력, 또는 선형예측 등을 이용하여 스펙트럼 영역에서 봉우리를 찾는 봉우리선택(peak-picking) 방법이 사용된다[10].



(그림 5) 스펙트럼 분포를 고려한 완만화

보간 포인트는 봉우리 선택 방법을 사용하여 포만트를 찾은 다음 그 주변의 스펙트럼 분포를 고려하여 결정된다. 보간을 수행하기 위해서는 두 음성신호의 연결부분에서 좌측과 우측에 있는 포만트들의 대응관계를 형성하여야 하는데, 순서가 서로 일치하도록 하여야한다. 즉, 좌측 1번 포만트와 우측 1번 포만트가 대응하게 되며, 2, 3, 4번 포만트에도 같은 방법으로 대응관계를 형성한다. 보간을 위한 포만트들의 대응관계가 정의된 다음에는 보간을 수행한 주파수 범위를 정의하는 보간 포인트를 산출한다. (그림 5)를 참조하여 이 과정을 설명하면 다음과 같다.

대응하는 두 개의 포만트를 f_0와 g_0라고 할때, 포만트를 중심으로 상하로 스펙트럼의 주파수 강도 값들을 적분한 결과가 단위 넓이 C가 되는 곳이 보간 포인트이다. 보간 포인트의 수는 포만트를 중심으로 상하 각각 m개 인데, 본 논문에서는 3 즉, 7개의 보간 포인트들을 사용하였다. 여기서, 단위 넓이 C는 2절에서 언급한 것처럼 불연속정도에 따른 인간의 청각신경 특성에 따라 적응적으로 결정된다.

각 보간 포인트는 식 (5)와 같이 (그림 5)에서 빗금 친 부분의 넓이가 A = A′와 B = B′가 되도록 결정된다. 식 (5)에서 $f(x)$와 $g(x)$는 두 음성신호의 주파수 강도 분포를 나타낸다. i는 포인트의 인덱스이다.

$$\int_{f_0}^{f_i} f(x)\,dx = \int_{g_0}^{g_i} g(x)\,dx = [\frac{i+1}{2}]C \tag{5}$$

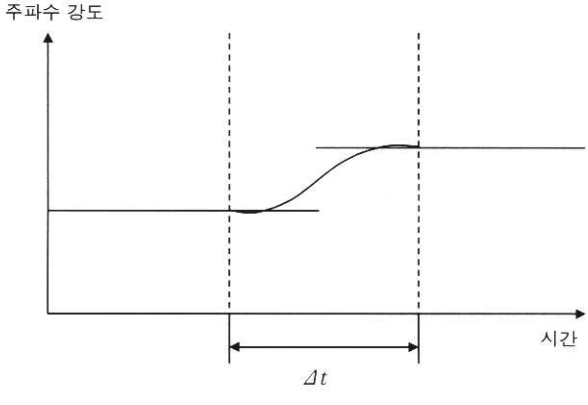
보간 포인트들이 결정된 다음에는 대응하는 포인트들에 의하여 형성되는 보간구간 Δt에서 스펙트럼 강도와 주파수 위치를 산출한다. 보간을 수행할 때는 완만화의 대상 구간인 Δt을 n개의 시간 위치로 나누어 시간축을 따라서 양쪽 스펙트럼의 주파수 강도 분포에 대한 가중치를 적용하여 식 (6)과 식 (7)과 같이 새로운 주파수 강도와 위치를 산출한다. 식 (6)과 식 (7)에서는 왜곡을 최소화하기 비선형적으로 보간을 수행한다. 식 (6)의 결과인 M은 비선형을 보간을 통하여 얻어진 새로운 주파수 강도이고, 식 (7)의 주파수의 새로운 위치이다.

$$M_n^j(k) = \frac{f_j^m \cdot NL \cdot (n-k) + g_j^m \cdot NL \cdot k}{n} \quad (0 \leq k \leq n,\ k \in Z) \tag{6}$$

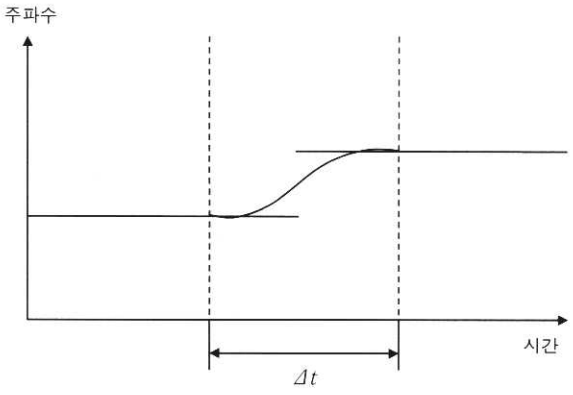
$$F_n^j(k) = \frac{f_j \cdot NL \cdot (n-k) + g_j \cdot NL \cdot k}{n} \quad (0 \leq k \leq n,\ k \in Z) \tag{7}$$

식 (6)과 식 (7)에서 j는 포인트의 인덱스이고 f_i^m은 j번째 포인트에서의 주파수의 강도이고 f_i는 그 포인트가 위치하는 주파수이다. NL은 비선형 함수인데, 본 논문에서는 식 8과 같이 B-스프라인(spline)을 사용한다[11]. 식 (7)에서 l의 범위는 구간길이에 따라 정해지며 $2l = \Delta t$가 되도록 정한다.

$$f(x) = \begin{cases} \frac{1}{2}|x|^3 - |x|^2 + \frac{2}{3} & 0 \leq |x| < l \\ -\frac{1}{6}|x|^3 + |x|^2 - 2|x| + \frac{4}{3} & l \leq |x| < 2l \\ 0 & 2l \leq |x| \end{cases} \tag{8}$$



(그림 6) 주파수 강도의 보간

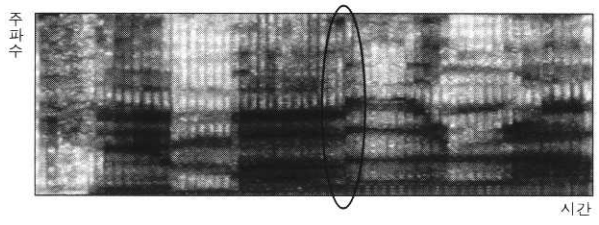


(그림 7) 주파수의 보간

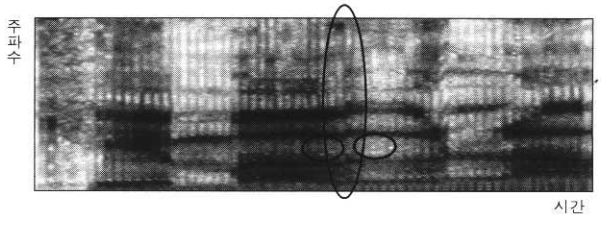
(그림 6)와 (그림 7)에서는 식 (6)와 식 (7) 그리고 식 (8)을 사용하여 스펙트럼 강도와 주파수 위치의 보간하는 과정을 가시화한 것이다.

4. 실험 및 결과

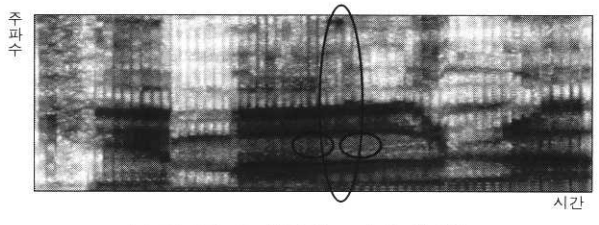
본 논문에서는 실험을 위해서는 ETRI 음성 DB 샘플과 5명이 녹음하여 제작한 총 20여개의 문장에서 추출한 약 500여 개의 다이폰을 사용하였다. 각 음성 샘플은 8KHz로 샘플링한 16bit 모노 음성 샘플이고 단위음소의 선택은 다이폰 군집화로 생성된 군집에서 추출한 다이폰을 사용하였다.



(a) 완만화 없이 연결한 경우



(b) 기존의 포만트 만을 고려한 완만화



(c) 포만트 및 분포를 고려한 완만화

(그림 8) 기존 완만화 방법과 제안한 방법의 스펙트로그램 비교

(그림 8)에서 음성합성의 예를 스펙트로그램(spectrogram)으로 보여준다. 이때, 사용된 단위음소는 '묵음 ㅣ'와 'ㅣ ㅐ'이다. (그림 8)(a)는 아무런 처리 없이 단위음소들을 단순히 연결한 경우 이고, (그림 8)(b)는 기존의 포만트만을 사용하여 완만화를 수행한 경우 이며, (그림 8)(c)는 제안된 방법으로 완만화를 수행한 경우이다. 스펙트로그램 상에서는 기존 방법과 제안된 방법 모두는 아무 처리를 하지 않은 경우 보다 자연스러운 스펙트럼의 연결을 보여주고 있다. 그러나, (그림 8)(b)에서 작은 두 개의 타원으로 표시된 양쪽의 각 부분의 밀도가 약간 다르게 나타남을 볼 수 있다. 이는 기존의 방법이 스펙트럼의 포만트에 초점을 맞추어 완만화를 수행하기 때문이다. (그림 8)(c)는 포만트 궤적 뿐 아니라 그 주변의 분포도 완만화 되었음을 볼 수가 있다.

<표 2>는 다이폰 군집화의 결과를 나타낸 표이다. 군집화는 LBG 알고리즘을 사용하여 모음 'ㅏ', 'ㅣ', 'ㅜ' 에 대해서 수행하였으며, 각 군집의 첫 번째 자음이 그 군집을 대표하는 자음이다.

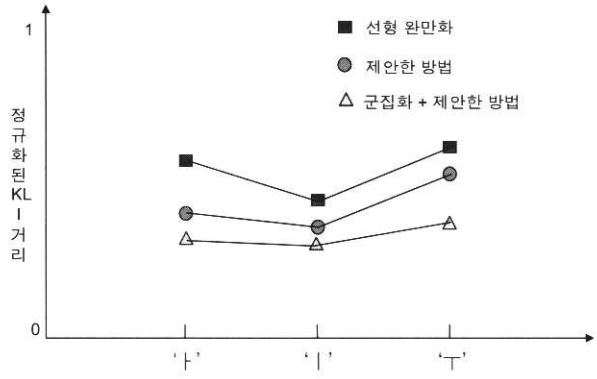
〈표 2〉 군집화 결과

모 음	군집에 속한 자음들
'ㅏ'	1 : ㅂ ㅍ ㅈ 2 : ㅅ 3 : ㅋ ㄱ ㄴ ㄷ ㄹ ㅁ ㅊ ㅌ ㅎ
'ㅣ'	1 : ㅋ ㄱ ㄴ ㅍ ㅌ 2 : ㅂ ㅈ ㄹ ㄷ ㅁ 3 : ㅅ ㅊ ㅎ
'ㅜ'	1 : ㄱ ㄴ ㅂ ㅋ 2 : ㄹ ㅈ ㅅ ㄷ ㅌ ㅊ 3 : ㅍ ㅎ ㅁ

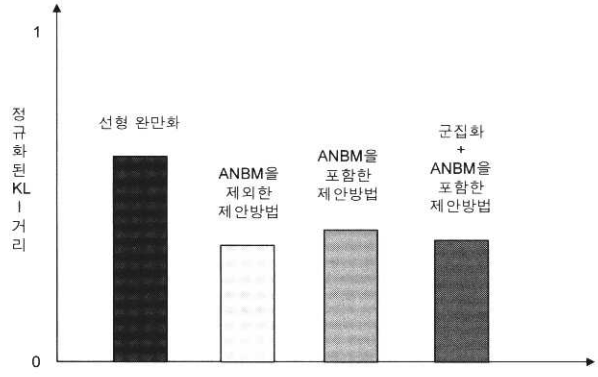
성능평가는 주관적인 관점과 객관적인 관점에서 수행하였다. 주관적인 관점에서의 성능평가를 위한 척도 MOS(Mean Opinion Score)을 사용하였고, 5명을 대상으로 합성한 문장들을 들려주어 합성된 음성이 듣기에 자연스러운 정도를 평가하였다[4]. 객관적인 관점에서의 성능평가는 단위음소의 연결 부분에서 식 (1)의 KL-거리를 사용하여 불연속 정도를 측정함에 의하여 이루어진다[6].

(그림 9)에서는 기존의 스펙트럼 완만화 방법과 제안된 방법 그리고 군집화된 다이폰을 단위음소로 사용한 제안 방법의 객관적인 관점에서의 성능의 비교를 보여준다. 이때 사용된 음소는 다이폰 기반의 음성합성에서 왜곡이 일어나기 가장 쉬운 음소인 'ㅏ', 'ㅣ', 'ㅜ'이다[7]. (그림 9)에 의하면 본 논문에서 제안된 방법이 보다 더 우수함을 알 수 있다.

(그림 10)은 같은 척도를 사용하여 기존의 방법과 비교하는 동시에 제안된 방법에서 ANBM의 적용 여부와 군집화 적용 여부에 따른 결과들을 비교하였다. 여기서 주목할 사실은 제안된 방법에 ANBM 점수를 적용한 경우와 적용하지 않은 경우를 비교했을 때, 오히려 ANBM 점수를 적용하지 않은 방법이 불연속을 더 줄이는 것으로 나타난다. 또한 ANBM을 적용한 방법에 군집화 방법을 적용하면 좀 더 불연속을 줄어드는 것으로 나타나고 있다.



(그림 9) 'ㅏ' 'ㅣ' 'ㅜ'의 음성합성 결과에 대한 KL-거리 비교



(그림 10) 3가지 방법에 대한 KL-거리 비교

KL-거리가 적다할지라고 음성합성을 통하여 산출된 결과 음성신호를 청취할 때 왜곡으로 인해 자연스럽지 못한 경우가 많다. 따라서, 합성된 결과 음성을 사람이 실제 청취한 후 주관적으로 성능을 평가하는 방법이 요구된다. 이를 위하여 본 논문에서는 5명을 대상으로 MOS를 측정하였다. 그 결과는 <표 3>과 같다.

〈표 3〉 MOS 테스트

알 고 리 즘	MOS
자연 음성	4.54
가공하지 않은 연결	3.21
기존의 선형적인 완만화	3.36
군집화 + 기존의 선형적인 완만화	3.50
ANBM을 사용하지 않은 제안방법	3.61
ANBM을 사용한 제안방법	3.82
군집화 + ANBM을 사용한 제안방법	4.02

(그림 10)과 <표 3>에 의하면, ANBM 점수를 적용하지 않은 방법이 객관적인 불연속 정도를 줄일 수는 있었으나

실제 사람이 듣는 합성음성의 질에서는 ANBM 점수를 적용한 방법이 더 나은 결과를 보여줌을 알 수 있다. 또한 여기에 군집화 방법을 적용할 경우 더 나은 결과를 보여주고 있다. 그 이유는 서로 다른 사람으로부터 추출된 단위음소를 사용하기 때문에 불연속이 더 크게 발생할 수 있는데, 이 경우 스펙트럼 완만화만을 수행했을때 생기는 왜곡을 군집화를 통하여 어느 정도 줄일 수 있기 때문이다.

5. 결론 및 향후 연구 과제

본 논문에서는 음성합성을 할 때 연결부위에서 생겨나는 스펙트럼의 불연속을 효과적으로 제거하는 방법을 제안하고 실험을 통한 성능평가를 수행하였다. 제안하는 방법은 단위음소 연결구간의 스펙트럼 분포특성의 유사도를 사용하여 단위음소들을 군집화하고 단위음소의 연결 구간에서 인간의 청각신경 특성을 고려하여 완만화의 양과 범위를 결정한 다음, 두 다이폰 경계의 스펙트럼 분포를 시간에 따라 가중치를 다르게 주어 스펙트럼 완만화를 수행한다. 성능평가는 객관적인 관점과 주관적인 관점에서 수행하였다. 그 결과 본 논문에서 제안한 방법이 기존의 주요 방법 보다 우수함을 알 수 있었다. 또한 군집화된 다이폰을 단위음소로 사용함으로써 음성 데이터베이스의 용량이 적은 경우나 서로 다른 화자로부터 녹음된 단위음소들을 가지고 있을 때 더 나은 성능을 기대할 수 있었으며, 추가 군집 생성시에 대표 자음을 중심으로 군집을 생성함으로써 데이터베이스의 크기를 작게 유지할 수 있었다.

본 논문에서 제안한 방법은 음성 데이터베이스를 구성함에 있어서 음성합성의 기본 단위로 사용될 단위음소가 얼마나 잘 추출되느냐에 크게 좌우된다. 따라서 단위음소 연결구간의 위상이나 음소의 길이에 따라 적절한 단위음소를 추출하는 방법에 관한 연구가 필요하며 이를 통하여 더욱 자연스러운 음성합성을 수행하는 방법에 대한 연구가 고려된다.

참 고 문 헌

[1] R. E. Donovan, P. C. Woodland, "A hidden Markov model based trainable speech synthesizer," Computer Speech and Language, pp.1-19, 1999.

[2] Conkie, A. D., Isard S., "Optimal coupling of diphones Progress in Speech Synthesis," Springer, New York, Chapter 23, pp.293-304, 1997.

[3] Kleijn W. B., Haagen J., "Waveform interpolation for coding and synthesis," Speech Coding and Synthesis, Chapter 5, pp.175-207, 1995.

[4] David T. Chappell, John H. L. Hansen, "A Comparison of Spectral Smoothing methods for segment concatenation based speech synthesis," Speech Communication Vol.36, pp.343-374, 2002.

[5] Wouters, J., Macon, M. W., "Control of Spectral Dynamics in Concatenative Speech Synthesis," IEEE Transactions on Speech and Audio Processing, Vol.9, No.1, pp.30-38, Jan., 2001.

[6] Esther Klabbers, Raymond Veldhuis, "Reducing Audible Spectral Discontivuities," IEEE Transactions on Speech and Audio Processing, Vol.9, No.1, Jan., 2001.

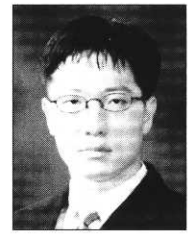
[7] H. van den Heuvel, B. Cranen, T. Rietveld, "Speaker variability in the coarticulation of /a, i, u/," Speech Communication, Vol.18, pp.113-130, 1996.

[8] Hossein Najafzadeh-Azghandi, "Perceptual Coding of Narrowband Signals," Ph.D Thesis, Department of Electrical & Computer Engineering, McGill University, Montreal, Canada, April, 2000.

[9] John H. L. Hansen and David T. Chappell, "An Auditory-Based Distortion Measure with Application to Concatenative Speech Synthesis," IEEE Transactions on Speech and Audio Processing, Vol.6, No.5, pp.489-495, Sep., 1998.

[10] L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signals," Prentice-hall, 1978.

[11] H. S. Hou and H. C. Andrews, "Cubic Splines for Image Interpolatio and Digital Filtering," IEEE Transactions on Acoustics Speech and Signal Processing, ASSP, Vol.26, No.6, pp.508-517, December, 1978.



장 효 종

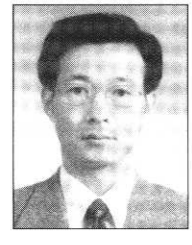
e-mail : ozjhj@vision.soongsil.ac.kr

2001년 숭실대학교 컴퓨터학부(공학사)

2003년 숭실대학교 대학원 컴퓨터학과 (공학석사)

2003년~현재 숭실대학교 대학원 컴퓨터학과 박사과정

관심분야 : 컴퓨터비전, 음성처리, 영상처리, 패턴인식, 3D 모델링 등



김 관 중

e-mail : kimkj@hanseo.ac.kr

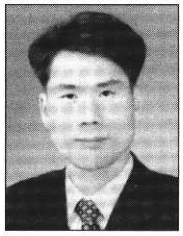
1983년 숭실대학교 전자계산학과(공학사)

1988년 숭실대학교대학원 컴퓨터학과 (공학석사)

1998년 숭실대학교대학원 컴퓨터학과 (공학박사)

1997년~현재 한서대학교 컴퓨터정보학과 조교수

관심분야 : 컴퓨터구조, 마이크로프로세서, 병렬처리, VLSI 설계 등



김 계 영

e-mail : gykim@computing.ssu.ac.kr

1990년 숭실대학교 전자계산학과(공학사)

1992년 숭실대학교대학원 컴퓨터학과 (공학석사)

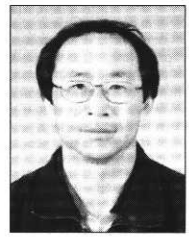
1996년 숭실대학교대학원 컴퓨터학과 (공학박사)

1996년~1997년 한국전자통신연구원(Post Doc.)

1997년~2001년 한국전력공사 전력연구원(선임연구원)

2001년~현재 숭실대학교 컴퓨터학부 조교수

관심분야 : 컴퓨터비전, 형태인식, 생체인식, 증강현실, 영상 및 신호처리 등



최 형 일

e-mail : hic@computing.ssu.ac.kr

1979년 연세대학교 전자공학과(공학사)

1982년 미시간대학교 전산공학과(공학석사)

1987년 미시간대학교 전산공학과(공학박사)

1987년~현재 숭실대학교 미디어학부 교수

관심분야 : 컴퓨터비전, 패턴인식, 퍼지이론, 비디오검색, 인터페이스 에이전트 등