개선된 스펙트럼 스무딩을 이용한 다이폰 클러스터링 기반의 연결 음성합성

장효종[0] 김계영 최형일
숭실대학교 컴퓨터학과
{ozjhj[0]}@vision.soongsil.ac.kr

{gykim, hic}@computing.soongsil.ac.kr

Concatenative Speech Synthesis based on Diphone Clustering using improved spectral smoothing

Hyo-Jong Jang[0] Gye-Young Kim Hyung-Il Choi
Dept. of Computing, Soongsil University

요 약

최근의 합성음성단위 연결을 통한 음성합성 방법의 잘 알려진 문제점은 연결 부분에서 불연속이 발생한다는 것이다. 본 논문에서는 음성을 합성할 때 나타나는 스펙트럼의 불연속을 제거하기 위하여 개선된 스펙트럼 스무딩 방법을 제안한다. 그리고 보다 좋은 스무딩의 결과를 얻기 위하여 음성합성의 단위로는 문맥에 민감한 클러스터링된 다이폰을 사용한다. 스무딩 방법에서는 연결 구간에서의 다이폰 바운더리에서의 양쪽 스펙트럼의 분포를 고려하여 시간에 따라 가중치를 다르게 주어 스무딩을 수행한다. 또한 가중치를 결정할 때 비선형 함수인 B-Spline함수를 사용하여 스무딩을 수행하여 보다 자연스러운 스펙트럼을 생성할 수 있었다.

1. 서 론

정보 통신의 발달로 인간은 컴퓨터를 이용하여 다양한 방법으로 정보를 교환하고 있다. 보다 신속하고 정확한 정보 교환을 위하여 인간과 컴퓨터 사이의 의사 교환 또한 중요한 이슈가 되고 있다. 이러한 관점에서, 음성은 인간의 가장 자연스러운 의사 전달 수단이며, 음성을 통한 컴퓨터와의 정보교환 기술은 매우 중요하다. 음성합성 시스템의 성능은 정확한 정보를 전달할 수 있도록 얼마나 자연스럽고 정확한 합성음을 만들 수 있는가에 달려있다. 자연스럽지 못한 합성음은 인간으로 하여금 이질감을 느끼게 하고 부정확한 합성음은 정확한 의사전달을 방해하게 된다. 현재 대부분의 음성합성 시스템은 음성 데이터베이스로부터 단위 음성 신호를 추출하고 이를 연결하여 합성음을 얻는다. 그러나 데이터베이스에 있는 합성 단위들은 모든 경우에 대한 문맥적인 차이나 변화들을 나타낼 수 없기 때문에, 이러한 합성 단위들을 가지고 합성음을 생성할 경우 연결 부분에서의 불연속이 발생하게 된다. 본 논문에서는 다이폰 클러스터링에 기반한 개선된 스펙트럼 스무딩 방법을 제안하여 이러한 문제를 해결하고자 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 관련연구 및 문제점에 관해서 살펴보고, 3장에서는 기존의 스펙트럼 스무딩 방법과 본 논문에서 제안하는 스펙트럼 스무딩 방법에 대하여 비교 설명한다. 4장에서는 실험 및 결과에 대해서 기술하며 5장에서는 본 논문의 결론을 맺는다.

2. 관련 연구

음성합성 단위의 연결 부분에서의 불연속을 해결하기 위한 기존의 연구에서는 다음과 같은 방법들이 사용되었다.

- 트라이폰과 같은 큰 합성 단위를 사용하는 방법[1]
- 스펙트럼 불연속이 최소화된 연결 위치를 찾아 이를 연결하여 불연속을 최소화하는 방법[2]
- 파형 스무딩, 스펙트럼 스무딩 등을 통해 스펙트럼의 불연속을 제거하는 방법[3]

이러한 세 가지 방법의 문제점을 살펴보면, 첫 번째 방법은 불연속이 완전히 없어지는 것이 아니라 빈도가 줄어들 뿐이다. 또한 큰 합성 단위를 쓰기 때문에 필요한 전체 데이터 양이 많고 데이터베이스의 크기도 증가하게 된다. 두 번째 방법은 연결 부분의 포만트 궤적이 수평이 아니라는 가정 하에 접근하는 방법이다. 이것은 두드러진 스펙트럼의 불연속이 나타나는 모음에서는 포만트의 궤적이 수평으로 나타나기 때문에 연결 위치의 변화만으로는 불연속을 제거할 수 없다. 세 번째 방법은 연결 부분의 불연속 정도에 따라서 부드러운 스펙트럼의 변화를 얻어내기 위해서 들이는 비용이 비싸며, 스무딩의 결과 자체도 신뢰하기 어렵다. 위와 같은 문제점을 해결하기 위한 대표적인 기존 방법들 중 하나가 다이폰 기반의 클러스터링 방법이다. 이 방법의 경우 음성을 합성할 때 두드러진 스펙트럼의 불연속이 나타나는 모음(/a/, /i/, /u/)에 초점을 맞추어 문제를 해결하고 있다.[4][5] 먼저 다이폰 연결 부분에서의 스펙트럼이 유사한 다이폰들을 클러스터링하여 적절한 크기의 데이터베이스를 구축한다. 이를 통해 트라이폰을 사용했을 때 현저히 늘어날 수 있는 데이터베이스의 크기를 줄일 수 있었고, 클러스터로부터 추출된 다이폰들을 합성하여 스펙트럼 불연속의 문제도 어느 정도 해결할 수 있다. 하지만 데이터베이스에 있는 합성 단위의 개수가 제한되어 있기 때문에 불연속을 완전히 제거하기가 쉽지 않다. 본 논문에서는 클러스터링된 다이폰을 기반으로 하고 여기에 개선된 스무딩 방법을 적용하여 보다 자연스러운 음성합성을 하고자 한다.

3. 스펙트럼 스무딩

3.1 기존 방법 및 문제점

기존의 방법에서는 합성 단위의 연결 부분에서 스펙트럼의 차이를 수정하여 불연속을 제거하는 여러 시도가 있었다. 그 중에 몇몇 방법에서는 스펙트럼 스무딩이 오히려 합성의 질을 떨어뜨리는 결과를 보여주기도 했다. 예를 들면 연결 부분에서 갑자기 나타나거나 혹은 사라지는 스펙트럼의 피크와 같은 스펙트럼의 왜곡이 생겨날 수 있다.[6] 이러한 단점을 해결하기

위해 데이터베이스로부터 이상적인 퓨전 유닛을 추출하여 이를 스펙트럼 스무딩에 이용하는 방법이 시도되었다.[7] 그러나 이 방법 또한 이상적인 퓨전 유닛을 추출하는데 드는 비용이 너무 크다는 것을 단점을 가진다. 본 논문에서는 클러스터링된 다이폰을 이용하여 스펙트럼 분포에 대한 스무딩을 시행한다. 클러스터링된 다이폰을 사용하는 이유는 데이터베이스로부터 추출된 다이폰을 사용하여 음성합성을 할 경우, 합성되는 다이폰 경계 사이의 유사도가 클수록 보다 좋은 스무딩의 결과를 얻을 수 있기 때문이다. 또한 기존의 스펙트럼의 궤적만을 주로 고려하는 스무딩 방법과 비교해 볼 때 전체적인 스펙트럼 분포를 고려한 스무딩 방법을 사용하여 보다 향상된 결과를 기대할 수 있다.

3.2 스펙트럼 분포를 고려한 스무딩

본 절에서는 본 논문에서 제안하는 스펙트럼의 포만트를 중심으로 그 주위 스펙트럼의 크기 분포를 고려하여 스무딩을 시행하는 방법에 대하여 설명한다. 그림 1은 전체 시스템 구조와 스펙트럼 스무딩의 세부 단계를 설명하고 있다. 먼저 데이터베이스에 있는 클러스터링된 다이폰들 중에 원하는 다이폰을 추출한다. 추출된 다이폰 경계 사이의 스무딩을 위해 보간 포인트를 탐색한다. 그 다음으로는 탐색된 포인트 위치에서의 스펙트럼의 크기를 계산한다. 마지막으로 계산된 스펙트럼의 크기가 위치하게 될 주파수를 계산한다. 보간을 할 때 접합구간과 대응되는 탐색 포인트사이의 거리를 n개의 구간으로 나누고 시간의 흐름에 따라서 양쪽의 스펙트럼 크기 분포에 대한 가중치를 조절하여 보간 한다. 여기서 기본 보간 방법은 B-Spline 방법을 사용한다. B-Spline 보간은 대응되는 탐색 포인트에서의 스펙트럼의 크기와 그 시점에서의 주파수의 위치를 알아내는데 사용하게 된다.

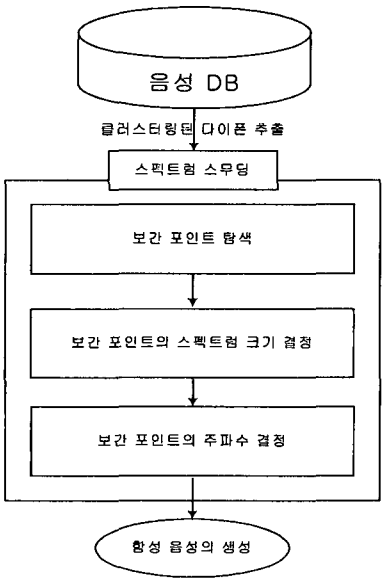

그림 1. 전체 시스템 구조

그림 2는 B-Spline 함수를 접합 구간과 탐색 포인트 사이에서 n개의 구간으로 나누어 적용하는 그림을 나타내고 있다. 기존의 선형적인 보간 방법과 비교해볼 때, B-Spline 함수를 이용하면 연결 부분에서의 스펙트럼 왜곡을 보다 더 줄일 수 있다. 또한 연결 부분에서의 스펙트럼의 분포를 고려하여 보간 포인트를 결정하기 때문에 연결 부분에서 노이즈처럼 갑자기 발생하는 스펙트럼의 왜곡 또한 방지할 수 있다.

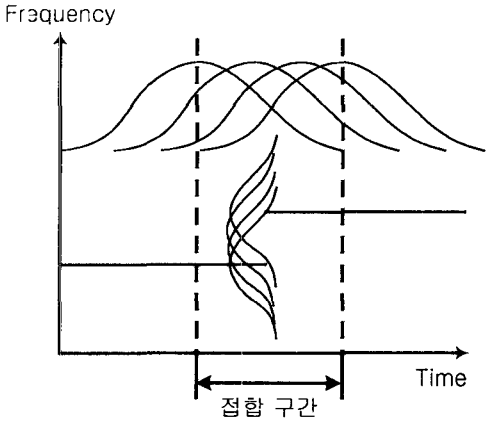

그림 2. B-Spline Interpolation을 이용한 스무딩

보간 포인트를 결정하기 위해서는 연결 부분의 포만트를 기준으로 좌우로 단위 배의 넓이만큼을 고려하여 보간 포인트를 결정한다. 이렇게 해줌으로써 포만트의 궤적뿐만 아니라 그 주변의 스펙트럼의 크기 분포도 반영이 된 스펙트럼 스무딩이 가능해 진다.

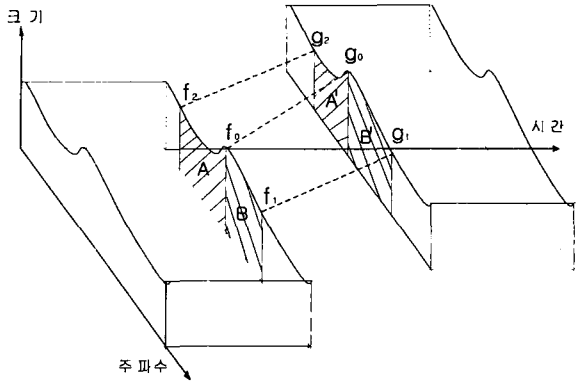

그림 3. 스펙트럼 분포의 스무딩

그림 3은 포만트를 중심으로 보간 포인트를 어떻게 탐색하는지를 나타내고 있다. 연결 부분에서의 좌우 포만트의 위치를 f_0, g_0라고 했을 때, 포만트 좌우의 스펙트럼의 크기들을 적분하여 그 넓이가 단위 넓이가 되는 곳이 구하고자 하는 보간 포인트이다. 즉, 그림 3에서 빗금 친 부분의 넓이는 A=A′, B=B′가 된다. 보간 포인트를 구하는 식은 다음과 같다.

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

식 (1)에서 [illegible]는 연결하는 다이폰 경계의 스펙트럼 크기의 분포를 나타낸다. i는 포인트의 인덱스이고 C는 단위 넓이를 나타내는 상수이다. 이렇게 보간 포인트가 구해지면 이 포인트를 기준으로 보간 한다. 다음에 나오는 두 수식에서 M

과 F는 스펙트럼의 크기와 포인트에서의 주파수 위치의 보간을 나타낸다.

$$M_n^j(k)=\frac{B(f_j)\cdot(n-k)+B(g_j)\cdot n}{n}(0\le k\le n,\ k\in Z) \quad (2)$$

$$F_n^j(k)=\frac{B(f_j)\cdot(n-k)+B(g_j)\cdot n}{n}(0\le k\le n,\ k\in Z) \quad (3)$$

기본적으로 보간 범위를 n개의 포인트로 나누어 시간의 흐름에 따라 가중치의 크기를 다르게 주도록 한다. j는 포인트의 인덱스이고 f_j는 그 포인트에서의 스펙트럼 크기이다. 이 크기의 가중치를 계산할 보간 함수는 다음과 같이 정의된다.

$$B(x)=\begin{cases}\frac{1}{2}|x|^3-|x|^2+\frac{2}{3} & 0\le|x|<l\\ -\frac{1}{2}|x|^3+|x|^2-2|x|+\frac{4}{3} & l\le|x|<2l\\ 0 & 2l<|x|\end{cases} \quad (4)$$

여기서 l는 접합구간과 주파수 보간 범위에 따라서 결정된다.

4. 실험 및 결과

본 논문의 실험을 위해서 사용한 음성 데이터베이스는 5명의 음성으로 녹음된 약 100여 개의 문장으로부터 추출된 다이폰들을 클러스터링하여 저장하고 있다. 실험은 주관적인 평가와 객관적인 평가로 나누어 진행하였다. 주관적인 평가로 MOS(Mean Opinion Score) 평가 방법을 사용하였고, 5명을 대상으로 합성한 문장들을 들려주어 합성된 음성이 얼마나 자연스러운가를 평가하였다.[7] 객관적인 평가로는 연결되는 다이폰 경계에서의 스펙트럼 분포끼리의 KL-Distance를 사용하여 얼마나 스펙트럼 스무딩이 잘 되었는지를 평가하였다.[4] 다음에 나오는 표 1과 그림 4는 각각의 테스트 결과이다.

Algorithm	MOS
Natural Speech	4.54
Raw concatenation	3.21
Waveform interpolation	3.12
Linear smoothing	3.36
Proposed smoothing	3.82

표 1. MOS TEST

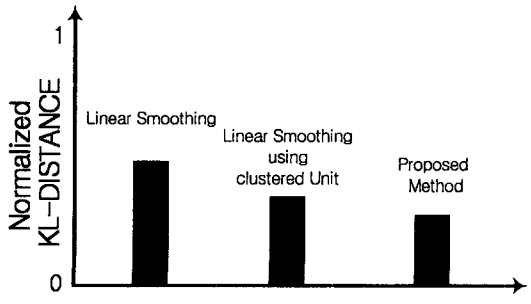

그림 4. 스무딩 방법에 따른 KL-DISTANCE 비교

5. 결론 및 향후 연구 과제

실험 결과에서도 알 수 있듯이, 본 논문에서 제안한 다이폰 경계의 분포를 고려한 스무딩 방법을 사용해 보다 향상된 합성음성을 생성할 수 있었다. 또한 연결 부분에서의 스펙트럼의 분포가 유사한 합성 단위를 이용하여 스펙트럼 스무딩의 결과를 더욱 향상 시켰다.

향후 연구과제로는 보다 유연하고 질 높은 합성음성을 생성하기 위해 발성 시 나타나는 억양, 강세, 리듬 등의 운율 정보를 스펙트럼 스무딩에 반영하는 것이다. 또한 연결 음성합성에서의 중요한 요소가 되는 합성 단위의 선정 문제와 이러한 합성 단위들을 어떻게 효과적으로 데이터베이스로 구축할 것인 가의 문제도 고려되어야 할 것이다.

Acknowledgement

본 논문은 첨단기술정보연구센터(AiTrc)를 통하여 과학재단의 일부를 지원 받았음.

참고 문헌

[1]. R.E. Donovan, P.C. Woodland, A hidden Markov model based trainable speech synthesizer, Computer Speech and Language, pp1-19, 1999

[2]. Conkie, A.D., Isard S., Optimal coupling of diphones Progress in Speech Synthesis, Springer, New York, Chapter 23, pp293-304, 1997

[3]. Kleijn W.B., Haagen J., Waveform interpolation for coding and synthesis, Speech Coding and Synthesis, Chapter 5, pp175-207, 1995

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

[5]. H. van den Heuvel, B.Cranen, T.Rietveld, Speaker variability in the coarticulation of /a,i,u/ , Speech Communication 18, pp113-130, 1996

[6]. David T. Chappell, John H.L. Hansen, A Comparison of Spectral Smoothing methods for segment concatenation based speech synthesis, Speech Communication 36, pp343-374, 2002

[7]. Wouters, J. ,Macon, M.W. ,Control of Spectral Dynamics in Concatenative Speech Synthesis, Speech and Audio Processing, IEEE Transactions on ,Vol 9 ,No. 1, pp30 -38, Jan 2001