코퍼스 기반 음성합성기에서 합성단위 선정을 위한 스펙트럼 거리측정 방법 비교연구

한승호* 김상진* 장경애** 한현배** 한민수*
* 한국정보통신대학교 음성/음향 정보 연구실
** Korea Telecom

Comparative study on methods for measurement of spectral distance for unit selection in corpus-based speech synthesis

Seung Ho Han*, Sang-Jin Kim*, Kyung Ae Jang**, Hyunbae Han**, and Minsoo Hahn*
* Speech and Audio Information Lab, Information and Communications Univ.
** Korea Telecom

space0128@icu.ac.kr

Abstract

Recently, most of speech synthesis systems are corpus-based concatenative ones using unit selection. Speech quality of these mainly depends on the unit selection from speech database. One of the important things in unit selection is a spectral distance measure. In this paper, we evaluated the speech quality of synthesized speech using different spectral distance measures. Implemented distances are the MFCC Euclidean distance, the log power spectral Euclidean distance, and the power spectral Kullback-Leibler distance. We performed the preference test on the output speech synthesized by the KT Hansori 2001 TTS synthesizer. Results show that the sentences using the power spectral Kullback-Leibler distance have the highest preference score.

Ⅰ. 서론

오늘날 많은 합성기들은 대용량 코퍼스 기반의 유닛 접합식 합성 방법을 사용한다. 이 방법은 음성 데이터베이스로부터 유닛(unit)을 가져와 이어주는 합성 방식이다. 대용량 음성 데이터베이스 안에는 우리가 원하는 유닛이 존재할 것이므로 보다 좋은 음질의 합성음을 생성할 수 있다는 것이 기본 개념이다. 원하는 유닛을 선택하는 과정을 unit selection이라고 부르며, 이 과정이 합성기 음질에 큰 영향을 미친다. 유닛을 선택하기 위한 방법으로 목표 비용(target cost)과 연결 비용(concatenation cost)의 합으로 이뤄진 비용 함수(cost function)가 제안되었다[1]. 목표 비용은 데이터베이스 안의 유닛과 목표 유닛이 얼마나 유사한지를 평가하는 것이고, 연결 비용은 연속적으로 이어지는 유닛들 사이의 연결 적합성을 평가해주는 비용이다. 특히, 연결 비용은 인간이 인지하는 청각적 불연속성(audible discontinuities)과 관련이 깊은 요소로 인식되고 있다[2]. 목표 비용과 연결 비용의 합인 전체 비용을 구한 후에는 Viterbi 검색을 통해 가장 적은 비용을 가지는 유닛 열을 찾아서 최종적으로 합성을 수행한다.

연결 비용을 계산 할 때 사용되는 특징 벡터들 중 하나가 유닛의 연결 지점에서의 spectral distortion이다. Spectral distortion을 측정하는 방법에는 여러 가지가 존재하며, 이러한 spectral distortion을 측정하는 스펙트럼 거리 측정 방법들 중 인간의 청각 특성과 유사한 스펙트럼 거리를 찾기 위한 연구가 수행되었다[2][3][4][5]. 특히, [2]에서는 고립단어 합성에 대해 총 13개의 스펙트럼 거리 측정방법에 대한 불연속성 검출 비교 연구를 수행하였다. 본 논문에서는 이 중 가장 성능이 우수하다고 알려진 3개의 스펙트럼 거리 측정 방법에 대해 문장 단위 합성으로 확장하여 비교 실험을 수행하였다. 또한, 불연속성 검출이 아닌 실제 합성음에 대한 선호도 평가 실험을 통해 연결 비용을 계산할 때 가장 적합한 스펙트럼 거리 측정 방법을 구한다. 본 실험에는 코퍼스 기반 유닛 접합식 합성기인 KT 한소리2001이 사용되었다[6].

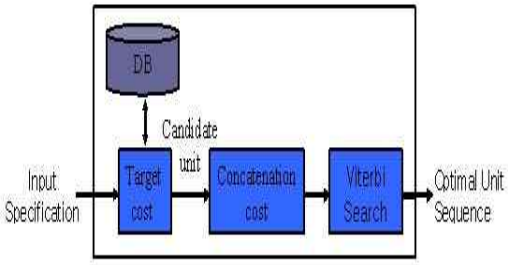

그림 1 Unit selection block diagram

Ⅱ. Unit Selection

실험을 위해 사용된 합성기의 unit selection 과정을 간단히 살펴보면 그림1과 같다. 먼저 unit selection 전단에서 구해진 입력에 대해 목표 비용을 계산하여 후보 유닛들을 선정한다. 이러한 후보 유닛들 중에서 다시 연결 비용을 계산하여 목표 비용과 가중치를 적용한 합을 통해 전체 비용을 구한다. 이것을 바탕으로 Viterbi 검색을 통해 최종 합성 유닛 열을 선정한다. 이러한 unit selection 과정을 거쳐 최종 합성 유닛 열을 선정하여 합성음을 생성한다[6].

1. Cost function

목표 비용의 계산은 피치(pitch), 세기(intensity), 발성 길이(duration), phone-environment, break index를 특징벡터로 사용하고, 연결 비용은 피치, 세기, spectral distortion을 특징벡터로 이용한다[6]. 이러한 특징벡터들 사이의 거리를 계산하여 가중치를 적용한 합을 구해 비용 함수로 사용한다. 본 실험에서는 스펙트럼 거리만의 영향을 고려하기 위해 연결 비용을 계산할 때 다른 특징벡터들은 배제하고 오직 스펙트럼 거리만을 이용하여 연결 비용을 계산하였다. 따라서 본 실험에서 사용되는 비용 함수는 식 (1)과 같다.

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n}\sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) + \sum_{i=2}^{n} w^c C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S) \quad (1)$$

여기서, $C_j^t(t_i, u_i)$는 목표 sub-cost이고, C^c는 연결 비용이다. n은 전체 유닛의 개수이고, p는 sub-cost의 수이다. 그리고 S는 묵음이며, u는 유닛이고, w는 가중치이다.

2. 스펙트럼 거리

스펙트럼 거리를 측정하는 여러 가지 방법 중 본 실험에서는 인간의 청각특성을 가장 잘 반영한다고 알려진 3개의 스펙트럼 거리 측정 방법을 이용하였다[2]. 스펙트럼 거리는 접합 유닛사이에서 측정되며, 본 실험에서 사용한 스펙트럼 거리 측정방법은 다음과 같다.

D1) MFCC 사이의 Euclidean 거리
D2) FFT 기반의 Log Power Spectra 사이의 Euclidean 거리
D3) FFT 기반의 Power Spectra사이의 Kullback-Leibler 거리

여기서, 두개의 특징 벡터 사이의 Euclidean거리와 Kullback-Leibler 거리는 각각 식 (2), (3)로 표현된다[2].

$$D_{Euclidean} = \sqrt{\sum_{n=1}^{p}(v_{1n} - v_{2n})^2} \quad (2)$$

$$D_{Kullback-Leibler} = \int (P(w) - Q(w)) log \frac{P(w)}{Q(w)} dw \quad (3)$$

Ⅲ. 실험 환경

1. 테스트 문장

실험에 사용된 문장은 표 1과 같다. 총 7개의 문장을 선정하여 사용하였다. 4개의 문장은 신문기사 중에서 랜덤하게 선정하였으며, 3개의 문장은 실제 음성 데이터베이스를 녹음할 때 사용한 문장들 중에서 랜덤하게 선정하였다.

2. 청취자

청취 실험을 위해 총 10명이 참여하였다. 그들은 모두 한국어를 모국어로 사용하는 사람들로 공학을 전공하는 대학원생들이다.

3. 청취환경

청취자는 실험실 환경에서 개인 PC를 통해 헤드폰으로 합성음을 듣고, 선호도 평가를 수행하였다. 합성음은 16비트 모노 음성 파일로 16 kHz의 표본화 주파수(sampling rate)를 가진다.

표 1 청취 테스트 문장

문장1	당시 현장에 있었던 필자는 아쉬워하는 많은 팬들의 표정을 직접 눈으로 볼 수 있었다.
문장2	아무쪼록 항상 그러했듯이 밝은 미소를 지니고 열심히 공부하는 지도자가 되기를 진심으로 당부한다.
문장3	현재 포항 이외에도 울산 현대와 FC 서울이 박주영에게 잔뜩 눈독을 들이고 있는 상태다.
문장4	잉글랜드 대표팀 주장 데이비드 베컴이 월드컵 예선에서 의도적으로 경고를 받았다고 고백, 구설에 시달리고 있다.
문장5	총풍 보고서 발견의 홍분과 함께 잠시 올라갔던 공안부 검사실 복도 앞의 철제 셔터도 다시 육중하게 내려졌다.
문장6	늙은 소나무, 낯익은 평나무와 보리둑나무 숲이 눈발 속에 자욱한 안개처럼 흐려 보일 뿐 조용한 산길이었습니다.
문장7	철천지 원수지간이었던 이츠하크 라빈 이스라엘총리와 야세르 아라파트 팔레스타인 해방기구의장의 악수는 어색했다.

Ⅳ. 실험

테스트 문장들에 대해 2절에서 설명한 3가지 종류의 스펙트럼 거리를 사용한 연결 비용 함수를 이용하여, 한 문장에 대해 각 3개씩의 합성음을 생성한다. 이 합성음에 대해 다음과 같이 총 3회의 선호도 평가를 실시한다.

T1) D1을 스펙트럼 거리로 사용하여 합성
vs. D2를 스펙트럼 거리로 사용하여 합성
T2) D2를 스펙트럼 거리로 사용하여 합성
vs. D3을 스펙트럼 거리로 사용하여 합성
T3) D3을 스펙트럼 거리로 사용하여 합성
vs. D1을 스펙트럼 거리로 사용하여 합성

선호도 평가결과를 바탕으로 제일 선호하는 문장에 3점, 두 번째 선호하는 문장에 2점, 세 번째 선호하는 문장에 1점을 부과한다. 만약, 청취자가 한 문장에 대해 올바르게 결과를 내리지 못했다면 그 결과는 제외시킨다. 예를 들어, T1 평가에서 D1을 사용한 합성음이 더 우수하다고 판별하고, T2 평가에서는 D2를 사용한 합성음이 더 우수하다고 판별하였는데, T3 평가에서 D3을 사용한 합성음이 우수하다고 판별하는 경우는 제외된다. 이와 같은 방법으로 모든 문장에 대해 평가하여 최종 점수를 산출한다.

Ⅴ. 결과

실험 결과 표2와 같이 D3을 스펙트럼 거리로 이용하여 합성한 문장들에서 가장 높은 선호도 점수를 얻었고, D1을 이용해 합성한 문장들이 두 번째로 높은 선호도 점수를 얻었고, D3을 이용해 합성한 문장들이 가장 낮은 선호도 점수를 얻었다.

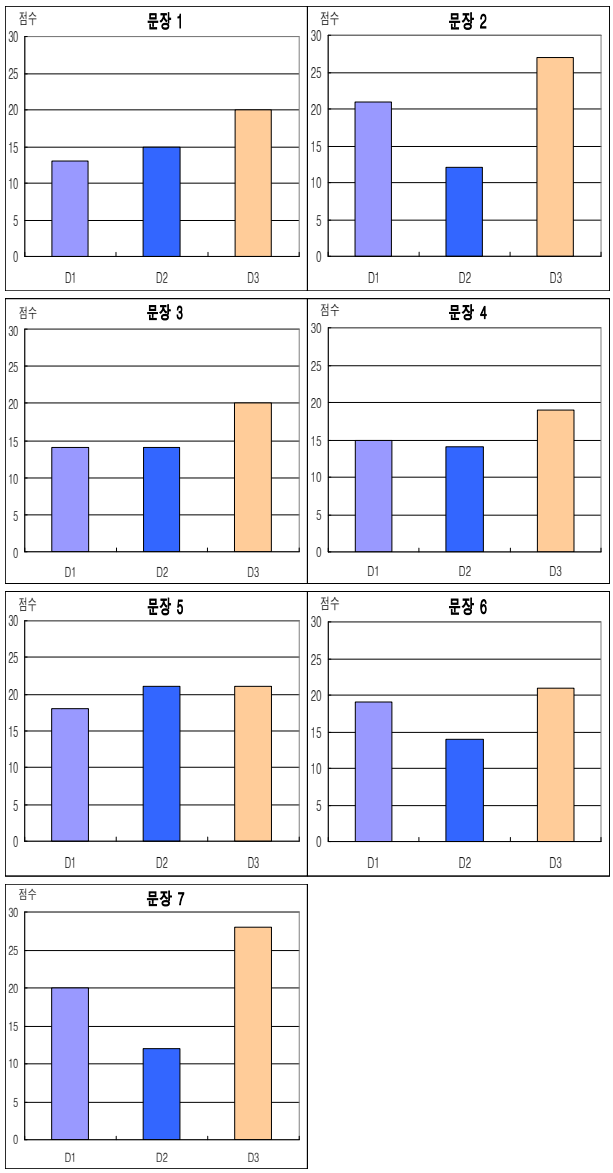

그림 2 선호도 조사 결과

표 2 선호도 조사 결과

	점수	백분율
D1	120	31.746%
D2	102	26.984%
D3	156	41.270%

각 문장들에 대한 청취도 실험 결과는 그림2와 같다. 모든 문장에서 D3을 스펙트럼 거리로 사용하여 합성한 경우가 가장 높은 점수를 보였다. 단, 문장 5에서는 D2, D3을 이용하여 합성한 경우에서 공동으로 가장 높은 점수를 보였다. D1이나 D2를 이용하여 합성한 경우에는 문장에 따라 그 결과가 다르게 나왔다. 문장 2, 4, 6, 7의 경우는 D1을 이용하여 합성한 경우가 D2를 이용해 합성한 경우 보다 높은 점수를 얻었고, 문장 1, 5의 경우는 D2를 이용하여 합성한 경우가 D1을 이용해 합성한 경우보다 높은 점수를 얻었다. 문장 3의 경우는 D1과 D2를 이용해 합성한 두 점수가 같게 나왔다.

이번 실험을 통해, D3이 인간의 청각 특성을 가장 잘 반영하는 것을 알 수 있다. D1은 전체적으로 고르게 인간의 청각 특성을 반영하지만 D3보다는 결과가 좋지 않았다. 그리고 D2의 경우는 합성하고자 하는 문장에 따라 그 편차가 크게 나는 것을 확인하였다. 또한, 기존의 고립단어 합성 실험을 통해 D3이 인간의 청각 특성을 가장 잘 반영하는 스펙트럼 거리라는 것이 알려졌는데, 이번 실험을 통해 실제 문장 단위 합성에서도 그러한 특성을 이용하여 스펙트럼 거리 측정 방법으로 활용하는 것이 바람직함을 확인하였다.

Ⅵ. 결론

우리는 unit selection 과정에서 사용되는 spectral distortion을 측정하는데 가장 적합한 스펙트럼 거리를 찾기 위한 실험을 진행하였다. FFT 기반의 power spectra 사이의 Kullback-Libler distance를 스펙트럼 거리로 사용하여 합성한 결과가 가장 우수한 결과를 보였다. 따라서 코퍼스 기반 유닛 접합식 합성에서 spectral distortion을 측정하기 위한 방법으로 위의 거리를 사용하는 것이 음질 향상에 도움을 줄 것이다. 또한, 하나의 스펙트럼 거리 측정 방법을 이용하는 것이 아닌, 우수한 여러 개의 스펙트럼 거리 측정 방법을 복합적으로 이용하여 사용해 볼 필요가 있으며 이를 연구 중이다.

참고문헌

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP*, 1996.

[2] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," *Proc. ICASSP*, 2001.

[3] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," *Proc. ICSLP*, 1998.

[4] J. Wouters and M. Macon, "Perceptual evaluation of distance measures for concatenative speech synthesis," *Proc. ICSLP*, 1998.

[5] R. Donoban, "A new distance measure for costing spectral discontinuities in concatenative speech synthesizer," The 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.

[6] A. Ferencz, S. Choi, H. Song, and M. Koo, "Hansori 2001-Corpus-based Implementation of the Korean Hansori Text-to-Speech Synthesizer," *Proc. EUROSPEECH,* 2001.

[7] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.