코퍼스 기반 음성합성기의 음질 향상을 위한 연결비용 함수에 관한 연구

한승호 [1], 김상진 [2], 장경애 [3], 구명완 [4], 한민수 [5]
한국정보통신대학교, 음성/음향 정보 연구실 [1 2 5]
KT 서비스 개발 연구소 음성언어연구팀 [3 4]
{space0128[1], sangjin[2], mshahn[5]}@icu.ac.kr, {kajang[3], mwkoo[4]}@kt.co.kr

A study on the concatenation cost function for improvement of speech quality in corpus-based speech synthesis

Seung Ho Han[1], Sang-Jin Kim[2], Kyung Ae Jang[3], Myoung-Wan Koo[4], and Minsoo Hahn[5]
Speech and Audio Information Lab., Information and Comm. Univ.[1 2 5]
Spoken Language Research Team, Service Development Lab., KT[3 4]

요 약

오늘날 많은 합성기들은 대용량 코퍼스 기반의 합성 방법을 사용한다. 이 방법은 음성 데이터베이스로부터 유닛(unit)을 가져와 이어주는 합성 방식이다. 합성 유닛을 선택하는 과정을 unit selection 이라고 부르며, 이 과정이 합성기 음질에 큰 영향을 미친다. 유닛을 선택하기 위해 목표비용(target cost)과 연결비용(concatenation cost)의 합으로 이뤄진 비용 함수가 사용된다. 각 비용은 특징 파라미터들 사이의 거리 비용에 가중치가 적용된 합으로 구할 수 있다. 본 논문에서는 합성음질의 향상을 위해 연결비용 계산과 관련된 방안을 제안한다. 첫째, 비용함수를 계산할 때 정규분포로 정규화된 특징 파라미터 값을 이용한다. 둘째, spectral distortion 을 측정할 때 power spectra 사이의 symmetric Kullback-Leibler 거리를 사용한다. 셋째, 음성 데이터베이스 내의 연속으로 녹음되어 저장된 유닛을 최종 합성 유닛 선정시 우선적으로 선택되도록 한다. 본 실험에는 코퍼스 기반 유닛 접합식 음성 합성기가 사용되었다. 제안된 알고리즘의 성능을 측정하기 위해 청취테스트를 수행하였으며, 제안한 방법을 적용한 결과 전체 청취자중 78%가 음질의 향상을 보고하였다.

Keyword: speech, synthesis, TTS, unit selection, corpus, concatenation cost, Kullback-Leibler distance

서 론

오늘날 많은 음성 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 selection 이라고 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)의 사용이 제안되었다[1,2]. 목표 비용은 데이터베이스 안의 유닛과 목표 유닛(target unit)이 얼마나 유사한지를 평가하는 것이고, 연결 비용은 연속적으로 이어지는 유닛들 사이의 연결 적합성을 평가해주는 비용이다. 목표 비용과 연결 비용

의 합인 전체 비용을 구한 후에는 Viterbi 검색을 통해 가장 적은 비용을 가지는 유닛열(unit sequence)을 찾아서 최종적으로 합성을 수행한다.

합성기의 음질은 예전에 비해 크게 개선되었지만 아직까지는 수요자들의 욕구에 완전히 충족시키지 못하고 있는 실정이다. 보다 자연스럽고 일관적인 합성음질을 얻기 위한 연구가 계속 진행되어야 한다. 따라서, 본 논문에서는 이를 위해 음성합성 시스템 중에서 unit selection 부분에 집중하여 음질을 향상 시킬 수 있는 방법에 대한 연구를 진행하였다. 특히, 어떠한 유닛이 연결되었는지는 음질에 상당한 영향을 미치는 요소이므로 보다 적절한 유닛이 연결될 수 있도록 연결비용 함수를 계산할 때 이용될 수 있는 방안을 제시한다.

본 론

1. 코퍼스 기반 음성합성 시스템

현재 주로 사용되는 음성합성 시스템은 코퍼스 기반 유닛 접합식 음성 합성 시스템이다. 이 합성방식은 1990 년대 중반에 발표된 논문에[1,2] 기초로 하고 있으며 이를 바탕으로 발전되어 왔다. 음성합성 시스템은 일반적으로 그림 1 과 같이 text specification 부, unit selection 부, 데이터베이스(DB), 그리고 신호처리(signal processing) 부로 나눌 수 있다. 사용자로부터 text 가 들어오면 시스템은 그 text 를 특성화(specification)하여 운율 예측, 단어사전에 없는 text 처리 등과 같은 자연어 처리를 한다[3]. 그리고, 구축된 DB 안에서 음성 유닛을 찾을 수 있도록 목표 특징값(target feature value) 등을 결정한다. 그러면 다음 단계에서 시스템은 주어진 목표 특징값과 가장 일치하는 유닛을 음성 데이터베이스 안에서 찾아 그것들을 이어주어 합성음을 생성한다. 유닛을 이어준 후 신호처리를 통해 음질을 개선시킬 수도 있으나 이상적인 코퍼스 기반 합성 시스템에서는 원하는 유닛이 반드시 존재함으로 후 신호처리과정은 필요없다[4].

본 논문에서 관심을 가지고 있는 unit selection 부에 대해 알아보면 다음과 같다. 연결할 유닛을

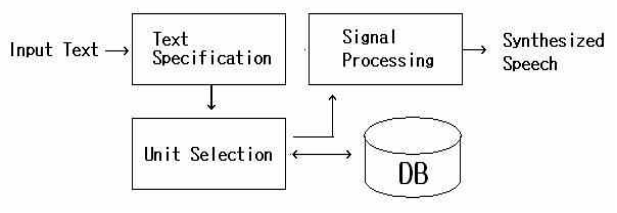

그림 1. Speech synthesis system

음성 데이터베이스로부터 선택하는 unit selection 과정은 합성음질에 미치는 영향이 매우 크기 때문에 중요한 과정이다. unit selection 과정은 다른 부분은 이상적이라는 가정 아래서 이뤄진다. 그림 2 에서와 같이, 먼저 unit selection 전단에서 구해진 입력에 대해 목표 비용을 계산하여 후보 유닛(candidate unit)을 선정한다. 이러한 후보 유닛들 중에서 다시 연결 비용을 계산하여 목표 비용과 가중치를 적용한 합(weighted sum)을 통해 전체 비용을 구한다. 전체 비용함수 식은 식 (1)과 같다[2]. 이것을 바탕으로 Viterbi 검색을 통해 최종 합성 유닛열을 선정하여 그것들을 연결시켜 합성음을 생성한다.

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n}\sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) + \sum_{i=2}^{n}\sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S_1) \quad (1)$$

여기서, $C_j^t(t_i, u_i)$, $C_j^c(u_{i-1}, u_i)$ 는 각각 목표 sub-cost 과 연결 sub-cost 이다. i 는 유닛 인덱스이고, j 는 sub-cost 인덱스이다. n 은 전체 유닛의 개수이고, p , q 는 sub-cost 의 수이다. 그리고 S 는 묵음이며, u 는 유닛이고, w 는 가중치이다.

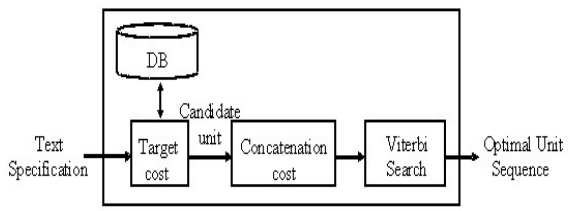

그림 2. Unit selection

2. 음질향상 방안

본 논문에서는 unit selection 과 관련되어 연결 비용함수 계산부분에 적용하여 음질을 향상 시킬 수 있는 방안을 연구한다. 보다 자연스러운 유닛의 연결을 통해 음질의 향상을 가져올 수 있다.

2-1 정규화된 특징 파라미터

연결 비용을 계산할 때 아래 식 (2)와 같이 정규분포로 정규화된 특징 파라미터를 비용을 계산한다. 연결 비용계산에 사용되는 여러 가지 종류의 특징 파라미터들은 서로 다른 특성을 가지며 특히, 다른 다이내믹 레인지(dynamic range)를 가진다. 예를 들어, 사용된 특징 파라미터가 피치와 세기라고 가정하자. 피치의 분포는 성인 여성의 경우 보통 200 Hz ~ 220 Hz 에 분포한다. 하지만 세기의 경우는 훨씬 넓은 분포를 가지며 그 값도 훨씬 크다. 그러므로 비록 가중치가 적용된다고 하더라도 이것들을 단순히 합하는 것에는 문제가 있다. 따라서 본 논문에서는 특징 파라미터 값을 정규분포로 변형시켜 다이내믹 레인지를 맞춘 후 가중치가 적용된 합으로 비용을 계산한다. 따라서, 각 특징 파라미터의 비용은 동등한 기여도로 표현된 상태에서 추가적인 가중치 최적화 과정을 거쳐 계산된다.

- Feature normalization: $$Z = \frac{x - \mu}{\sigma} \qquad (2)$$

여기서, x 는 특징 파라미터 값이고, μ 는 그들의 평균이며, σ 는 표준분포이다. 따라서, Z 는 정규화된 특징 파라미터 값이다.

2-2 Spectral distortion 계산

연결되는 유닛 사이에 spectral distortion 이 발생하면 사람들은 음질의 저하를 크게 느낀다. 그러므로, 연결 비용을 계산할 때 spectral distortion 은 하나의 특징 파라미터로써 사용되며, 인간의 청각특성과 유사한 spectral distortion 을 측정하기 위한 방법에 대한 연구가 진행되고있다[5,6,7,8]. 여러 방법 중 power spectra 사이의 symmetric Kullback-Leibler 거리가 인간의 청각특성과 가장 유사하다고 알려졌다. 이를 확인하기 위한 선행연구로써 실제 합성기에 적용하여 거리 측정 방법에 따른 합성음의 음질을 비교 평가한 실험에서도 역시 symmetric Kullback-Leibler 거리를 이용하는 것이 합성음질의 향상을 이룰 수 있다는 것을 알았다[9]. 따라서, 본 논문에서는 음질향상을 위해 이 거리측정방법을 실제 합성기에 적용시켜 다른 특징 파라미터들과 함께 사용한다. symmetric Kullback-Leibler 거리는 식 (3)과 같이 정의된다 [10].

$$D = \int (P(w) - Q(w)) \log \frac{P(w)}{Q(w)} dw \qquad (3)$$

2-3 원 데이터베이스에서 연속된 유닛의 연결

연결되는 유닛 사이에 연결오차가 크면 합성음질이 나빠진다. 이러한 오차를 줄이기 위해 합성 시스템은 연결 비용 계산을 통해 가장 적은 연결비용을 가지는 유닛을 연결한다. 그런데, 만약 합성하려는 문장 중에 원 음성 데이터베이스 안에 연속으로 존재하는 유닛이 연결된다면, 그 유닛 사이의 연결 오차는 적을 수 밖에 없다. 하지만 시스템이 특징 파라미터들 사이의 연결 비용을 계산하여 합성 유닛을 선정하면 그런 유닛이 선정되지 않을 수도 있다. 따라서, 합성 유닛을 선정할 때 후보 유닛 중 그러한 유닛이 있다면 인위적으로 그 유닛이 최종 합성 유닛으로 선정되게 한다. 물론, Viterbi 탐색 과정에서 위와 같이 연결된 유닛이 포함된 합성열이 선택되지 않는다면 그 효과는 나타나지 않을 수 도 있다.

3. 실험

제안된 알고리즘의 성능을 평가하기 위해 주관적인 선호도 조사가 실시되었다. Baseline 합성기의 합성음과 제안된 방법을 적용하여 합성한 합성음 사이의 상대적인 음질을 비교하였다. 객관적으로 합성 음질을 비교 평가하기위해 일반적으로 사용되는 방법이 존재하지 않는다. 따라서, 주관적인

음질 평가로써 합성음의 음질을 비교하여 제안된 알고리즘의 성능을 평가하였다.

3-1 Baseline 합성기

합성기의 음질을 개선 실험시 사용된 baseline 합성기는 대용량 코퍼스 기반의 유닛 접합식 음성 합성기이다[11]. Unit selection 의 연결비용 함수에 사용된 정보는 cepstral distortion, pitch, energy 이며 합성단위는 triphone 이다.

3-2 청취실험환경

합성 문장은 총 8 개의 문장으로 구성되어 있다. 객관적인 합성문장 선정을 위해 4 개의 문장은 신문기사에서, 나머지 4 개 문장은 실제 음성 데이터베이스가 녹음될 때 사용된 문장에서 무작위로 선정했다. 합성음은 16 kHz 의 표본화율(sampling rate)을 가지며, 16 비트의 모노 웨이브 파일이다. 청취는 실험실 환경에서 개인 PC 를 가지고 헤드폰으로 이루어졌다. 청취 실험에 총 18 명이 참여했으며 모두 한국어를 모국어로 사용하며 공학을 전공하는 학생들이다.

3-3 청취실험방법

선호도 청취 테스트는 CCR(Comparison Category Rating) 방법[12]으로 알려진 방식으로 이뤄졌다. Baseline 합성기와 제안된 방법을 적용한 합성기를 이용하여 각 테스트 문장에 대해 각각의 합성음을 생성한다. 2 가지 합성음에 대해 한 문장을 기준으로 나머지 문장의 음질을 상대적으로 –2, -1,0,1,2 으로 5 단계로 평가한다. 기준 문장은 각 문장 청취 시 마다 무작위로 선택된다. 총 8 문장에 대한 평가를 마친 후 baseline 합성기의 합성음을 기준으로 다시 정리하여 결과를 산출한다.

4. 결과

청취 선호도 조사 결과 제안된 방법을 적용하여 합성한 합성음의 음질이 기존 방법에 의한 합성음의 음질보다 전반적으로 우수한 것으로 나타났다. 합성 문장, 청취자에 따라 물론 선호도 정도, 선호하는 음성은 차이를 보였다.

표 1. 문장별 청취테스트 결과

	-2	-1	0	1	2	합	평균	표준편차
문장 1	0	4	1	13	0	9	0.50	0.86
문장 2	0	2	6	10	0	8	0.44	0.70
문장 3	2	3	4	8	1	3	0.17	1.15
문장 4	0	0	4	7	7	21	1.17	0.79
문장 5	0	1	7	10	0	9	0.50	0.62
문장 6	0	5	6	6	1	3	0.17	0.92
문장 7	0	1	7	9	1	10	0.56	0.70
문장 8	0	4	7	5	2	5	0.28	0.96
평균						8.50	0.47	0.84

4-1 음질향상

표 1 과 같이 테스트 문장 별로 분석해 보면 문장 1,4,5,7 의 경우는 평균이상의 선호도 점수를 획득하였다. 즉, 비교적 음질 향상의 정도가 비교적 크다고 볼 수 있다. 나머지 문장인 2,3,6,8 의 경우는 선호도 점수 평균이 평균 이하이다. 따라서, 음질의 향상이 발생은 하였지만, 비교적 크지 않음을 유추할 수 있다. 문장 1,3,6,8 의 경우는 선호도 점수의 표준편차가 평균이상을 보임을 확인할 수 있다. 즉, 청취자에 따라 음질 향상 또는 저하의 정도가 크게 차이 난다고 말할 수 있다. 그리고 문장 2,4,5,7 의 경우는 선호도 점수의 표준편차가 평균 이하 값을 보이므로 청취자 대부분이 비슷하게 음질을 평가했음을 알 수 있다.

표 2 와 같이 전체 청취테스트 결과를 보면, 총 18 명의 청취자 중 14 명의 청취자는 제안된 방법을 적용하여 합성한 음성에 대해 전반적으로 음질이 향상된 것으로 느끼는 것으로 나타났다. 2 명의 청취자는 전반적으로 음질의 변화가 없는 것으

표 2. 전체 청취테스트 결과

음질변화	선호도 점수 평균	수
음질향상	0.5 이상	10 명
	0~0.5	4 명
변화없음	0	2 명
음질저하	-0.5~0	2 명
	-0.5 이하	0 명
총 합		18 명

로 느꼈으며, 2 명의 청취자는 음질이 조금 나빠졌다고 응답하였다. 음질이 나빠졌다고 느낀 청취자에 경우도 그 정도는 그렇게 크지 않았다. 음질이 좋아졌다고 느낀 청취자 중 4 명은 그 느끼는 정도가 평균 0.5 이하였고, 나머지 10 명은 평균 0.5 이상의 상당히 큰 음질 향상을 느꼈다. 정리하면, 전체 청취자중 78%가 음질이 나아졌다고 느꼈으며, 11%가 음질의 변화가 거의 없다고 느꼈고, 단지 11%가 음질이 나빠졌다고 느꼈다. 따라서, 테스트 결과 대부분의 사람이 음질의 향상을 느꼈음을 알 수 있다.

4-2 Unit 변화

제안한 알고리즘을 적용함으로써 최종 유닛 선정시 후보 유닛중 약 77%정도가 다른 유닛으로 변경되었다. 제안한 방법 중 첫번째와 두번째 방법은 Viterbi 탐색에 전체 경로에 영향을 주며, 세번째 방법은 부분적으로 영향을 줄 수 있다. 그러나, 모든 유닛 변화가 음질에 영향을 주는 것은 아니며 청취자가 음질을 듣고 판단하는 방법 이외에 다른 방법으로 유닛 변화로 인한 영향을 판단하는 것은 힘들다. 그리고 음질의 차이는 음성의 모든 부분에서 느껴지는 것이 아니라 음성 일부분에서 느껴진다. 따라서, 제안된 방법에 의해 변화된 모든 합성 유닛의 변화가 음질에 영향을 주었다기 보다, 어느 특정 부분의 유닛 변화가 음질 판단에 영향을 미쳤다고 볼 수 있다.

결 론

코퍼스 기반 합성기에서 unit selection 과정은 매우 중요한 부분이다. 어떠한 유닛이 선택되고 이어지는 가에 따라 음질이 크게 달라지기 때문이다. Unit selection 과정은 크게 목표 유닛에 적합한 유닛을 찾는 과정과 그렇게 찾은 유닛들을 적절하게 이어주는 과정으로 분리할 수 있다. 물론, 두 과정 모두가 중요하겠지만, 인간은 합성음을 평가할 때 얼마나 적절하게 유닛이 연결되었는 지에 좀 더 민감하다.

따라서, 본 논문에서는 음질의 향상을 위해 연결비용 계산 부분에 집중하여 보다 정확한 유닛이 연결 될 수 있는 방법에 대해 제안을 하였다. 첫째, 특징 파라미터들을 정규분포로 정규화하여 사용한다. 둘째, power spectra 사이의 symmetric Kullback-Leibler distance 를 spectral distortion 측정에 이용하였다. 셋째, 음성 데이터 정보를 이용하여 원 데이터베이스에서 연속으로 존재하는 유닛을 최종 합성 유닛으로 선택한다.

제안된 알고리즘의 성능 평가를 위해 기존의 합성기에서 합성한 음과 제안된 알고리즘을 적용하여 합성한 음에 대한 선호도 조사를 실시하였다. 그 결과 제안된 알고리즘을 적용하여 합성음을 생성한 합성음의 음질이 기존의 방법으로 합성한 경우 보다 크게 향상됨을 알 수 있다. 비록 청취자와 문장에 따라 그 정도의 차이가 존재하지만, 모든 문장에서 일반적으로 음질이 향상된 것으로 조사되었으며, 총 18 명의 청취자중 78%인 14 명의 청취자가 음질의 향상을 느꼈다. 11%인 2 명은 비슷하다고 응답하였다. 그리고 단지 11%인 2 명만이 음질의 저하를 보고하였다. 따라서, 제안한 방법을 적용하면 음질의 향상을 가져올 수 있음을 알 수 있다.

감사의 글

본 연구는 KT 와 한국정보통신대학교 디지털 미디어 연구소의 프로젝트에 의하여 지원되었음.

참고문헌

[1] A. W. Black, N. Campbel, “Optimising Selection of Units from Speech Databases for Concatenative Synthesis,” *Proc. of Eurospecch95,* 1995.

[2] A. J. Hunt, A. W. Black, “Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database,” *Proc. of ICASSP96*, 1996.

[3] J. Schroeter, A. Conkie, et al., “A Perspective on the Next Challenges for TTS Research,” *Proc. of IEEE Workshop on Speech Synthesis*, 2002.

[4] A.W. Black, “Perfect Synthesis for All of the People

All of the Time," *Proc. of IEEE Workshop on Speech Synthesis*, 2002.

[5] Y. Stylianou, A. K. Syrdal, "Perceptual and Objective Detection of Discontinuities in Concatenative Speech Synthesis," *Proc. of ICASSP2001*, 2001.

[6] J. Vepa, S. King, P. Taylor, "New Objective Distance Measures for Spectral Discontinuities in Concatenative Speech Synthesis," *Proc. of IEEE Workshop on Speech Synthesis*, 2002.

[7] J. Wouters, M. Macon, "Perceptual Evaluation of Distance Measures for Concatenative Speech Synthesis," *Proc. of ICSLP98*, 1998.

[8] 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.

[9] 한승호 외 4 명, "코퍼스 기반 음성합성기에서 합성단위 선정을 위한 스펙트럼 거리측정 방법 비교연구," 대한음성학회 가을학술대회, 2004.

[10] R. Veldhuis , E. Klabbers, "On the Computation of the Kullback-Leibler Measure for Spectral Distances," *Proc. of IEEE Transactions on Speech and Audio Processing*, 2003

[11] A. Ferencz, et al., "Hansori2001-Corpus-based Implementation of the Korean Hansori Text-to-Speech Synthesizer," *Proc. of Eurospeech2001*, 2001.

[12] X. Huang, A. Acero, H. Hon, *Spoken Language Processing*, Prentice Hall, 2001.