코퍼스 기반 음성합성기를 위한 합성단위 경계 스펙트럼 평탄화 알고리즘

김상진(ICU), 장경애(KT), 한민수(ICU)

<Contents>

<Abstract>

A Spectral Smoothing Algorithm for Unit Concatenating Speech Synthesis

Sang-Jin Kim, Kyung Ae Jang, Minsoo Hahn

Speech unit concatenation with a large database is presently the most popular method for speech synthesis. In this approach, the mismatches at the unit boundaries are unavoidable and become one of the reasons for quality degradation. This paper proposes an algorithm to reduce undesired discontinuities between the subsequent units. Optimal matching points are calculated in two steps. Firstly, the Kullback-Leibler distance measurement is utilized for the spectral matching, then the unit sliding and the overlap windowing are used for the waveform matching. The proposed algorithm is implemented for the corpus-based unit concatenating Korean text-to-speech system that has an automatically labeled database. Experimental results show that our algorithm is fairly better than the raw concatenation or the overlap smoothing method.

* Keywords: Spectral smoothing, Spectral mismatch, Unit concatenation, Kullback-Leibler distance, Speech synthesis, Spectral matching, Waveform matching, Overlap windowing.

1. 서 론

코퍼스 기반의 합성단위 연결 음성합성 시스템은 대용량 음성 데이터베이스로부터 적절한 음성 합성단위를 선택하고 연결하여 음성을 합성한다[1]. 이 방법은 뛰어난 음질의 결과물을 얻을 수 있기 때문에 현재 가장 널리 쓰이고 있다. 그러나 이 기술은 음소 분할된 음성 데이터베이스에 크게 영향을 많이 받는다. 즉, 데이터베이스로부터 가장 적절한 합성단위를 선택하여 음성을 합성하기 때문에, 만약 목표 합성단위들에 적합한 이상적인 합성단위들이 모두 존재하도록 데이터베이스가 충분히 크다면 이 시스템은 이상적인 음성을 합성할 것이다. 그러나 일반적으로 데이터베이스의 크기는 한계가 있다. 게다가, 입력 문장이 데이터베이스에 포함되어있지 않은 합성단위들을 포함하고 있을 때, 그런 합성단위들의 대체 합성단위들은 연결되는 경계에서 불일치를 야기 시킨다[2]. 또한 데이터베이스의 크기가 충분히 크더라도 합성단위 사이의 불연속성은 여전히 발생할 수 있고, 결과적으로 합성 음성의 질을 떨어뜨릴 수 있다. 따라서 이러한 원치 않는 불일치를 줄이기 위해 스펙트럼의 평탄화가 필요하다. 본 논문에서는 합성단위 경계의 스펙트럼 평탄화 방법을 제안하며 실험을 통해서 그 유용성을 확인하였다.

본 논문의 구성은 다음과 같다. 먼저, 2절에서는 스펙트럼 평탄화에 대해 소개하고, 제안된 알고리즘을 설명하였다. 3절에서 실험 및 결과를 제시하였으며, 4절에서 결론을 맺었다.

2. 스펙트럼 평탄화

만약 음성 데이터베이스가 합성시 필요한 목표 합성단위들을 모두 수용할 만큼 충분히 크고, 각 음성 합성단위들이 매우 잘 분할되어 있다면, 각 합성단위들을 별다른 신호처리 없이 연결해도 좋은 음질의 합성음성을 생성할 수 있을 것이다. 그러나 후보 합성단위의 숫자는 제한되기 때문에 합성단위 연결을 통한 음성합성에서 합성단위가 천이되는 구간에서의 스펙트럼의 불연속은 빈번하게 발생한다. 만일 적절한 스펙트럼 평탄화 기술이 적용된다면, 합성단위들의 경계에서 발생하는 불일치를 줄일 수 있을 것이다. 이미 몇몇 연구 논문들에서 스펙트럼 평탄화 기술을 소개했지만 그 기술들은 대부분 음색 변환이나 운율 변경을 위해서 발전되고 적용되어 왔다[3][4].

Chappell과 Hansen은 기존의 평탄화 알고리즘들을 구현하고 실험하여 MOS 결과를 제시하였다[3]. 실험된 알고리즘들은, raw concatenation, optimal coupling, waveform interpolation, the LP technique (pole shifting and LSF interpolation), 그리고 the continuity effect 기술이다. 그들의 실험 결과에 따르면, 단지 optimal coupling

알고리즘만이 raw concatenation보다 좋은 결과를 보인다. 이것은 다른 알고리즘들은 대개의 경우 연결되어진 합성단위들의 음질을 떨어뜨린다는 것을 의미한다. 또한, Chappell과 Hansen은 그들의 논문에서 스펙트럼 평탄화 기술은 충분히 크거나 혹은 특별히 디자인된 데이터베이스에 적용시 더욱 효과적임을 언급하였다[3]. 본 논문에서는 그들의 실험에서 raw concatenation보다 좋은 결과를 보여준 유일한 알고리즘인 optimal coupling 알고리즘만 고려하였다.

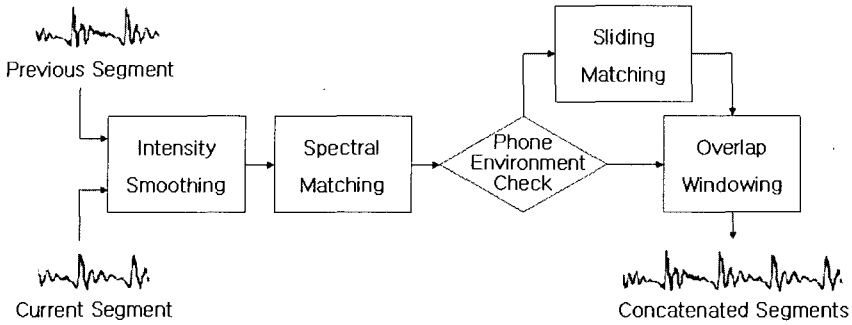
2.1. Optimal Coupling 알고리즘

Conkie와 Isard가 제안한 optimal coupling 알고리즘에서 최적의 연결지점은 합성단위들 간의 객관적인 스펙트럼 거리측정을 통해서 구한다[5]. 그들이 사용했던 거리측정 방법은 MFCC 특징벡터 기반의 유클리디언 거리측정이다. Optimal coupling 알고리즘은 구현이 간단하다는 장점을 지닌다. 게다가 이 알고리즘은 합성단위를 변경하지 않기 때문에 합성단위의 음질저하를 동반하지 않는다. 그러나 최적의 연결지점을 찾기 위한 연산량이 많으며 시간이 많이 걸린다는 단점을 지닌다.

2.2. 제안된 Spectral Smoothing 알고리즘

제안된 스펙트럼 평탄화 알고리즘은 optimal coupling 알고리즘을 기반으로 고안되었으며, 최적의 연결지점을 검색하는 시간을 줄이기 위해서 미리 정해진 일치 영역에서만 연산을 한다. 제안된 알고리즘은 <그림 1>과 같으며 다음과 같은 단계로 세분화 할 수 있다:

- Intensity smoothing,
- Spectral matching,
- Phone environment checking,
- Waveform matching,
- Overlap windowing.



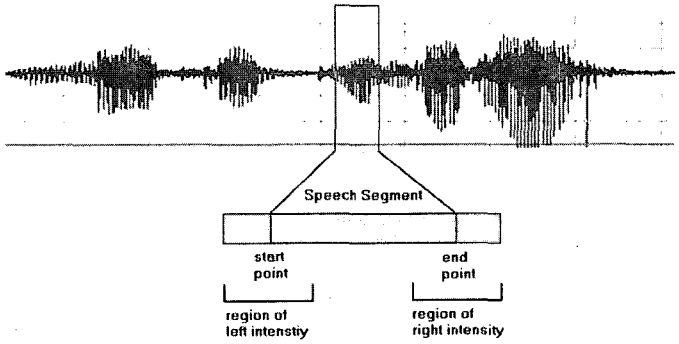
<그림 1> 제안된 스펙트럼 평탄화 알고리즘

2.2.1. Intensity Smoothing

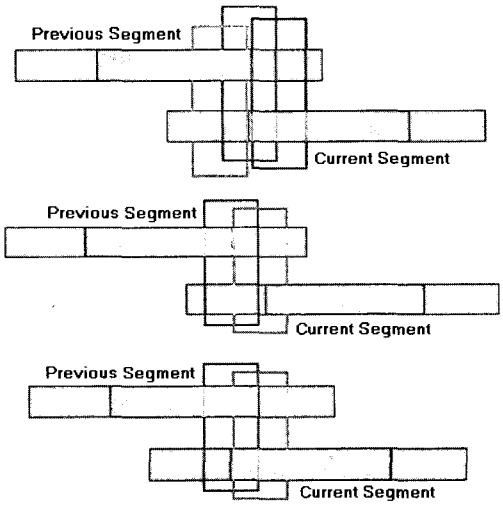
음성 합성을 위해 선택된 합성단위들은 각기 다른 음성으로부터 얻어지기 때문에 합성단위 경계에서 각 합성단위들의 에너지는 차이를 보인다. 비록 합성단위 선택 알고리즘의 연결 비용 함수에서 에너지 차이를 고려하지만, 합성단위의 연결단에서 에너지 정규화를 이용하면 더욱 효과적으로 각 합성단위를 연결할 수 있다. 먼저, 식(1)에서와 같이 이전 합성단위의 오른쪽 에너지와 현재 합성단위의 왼쪽 에너지 사이의 상관관계를 계산한다.

$$\gamma_i = \frac{Unit(i-1)_{right}^{intensity}}{\sqrt{Unit(i-1)_{right}^{intensity} \times Unit(i)_{left}^{intensity}}} \tag{1}$$

여기서, $Unit(i)^{intensity}$ 는 합성단위로부터 구한 에너지를 의미한다. 음성 데이터베이스의 모든 합성단위들은 <그림 2>에서 볼 수 있듯이 음성 합성단위의 시작부분과 끝부분 양쪽 모두 바깥부분에 여분의 음성을 포함하도록 분할하였다. 합성단위의 왼쪽 및 오른쪽 에너지, 즉 $Unit(i)_{left}^{intensity}$과 $Unit(i)_{right}^{intensity}$은 이러한 여분의 음성을 포함한 일부 구간에서 구하였다. 식(1)을 통해 구한 γ_i 값은 합성단위 경계 부분의 에너지 비율이며, 그 값이 0보다 크면 다음 합성단위의 에너지를 γ_i 값의 비율만큼 증가시켜야 함을 의미하며, 작으면 감소시켜야 함을 반영한다.



<그림 2> 여유 정보를 포함한 합성 합성단위의 예



<그림 3> 스펙트럼 거리측정을 위한 후보 7쌍의 예

2.2.2. Spectral Matching

Klabbers와 Veldhuis는 SKL (symmetric Kullback-Leibler) 거리측정 방법이 스펙트럼의 불연속성 측정에 효과적임을 제시하였다[6][7]. 인간이 인지할 수 있는 음성의 불연속성은 대부분 주파수 영역에서의 갑작스러운 변화로부터 온다. 따라서 연결 합성단위 사이의 스펙트럼 불일치가 가장 적은 최적의 접합 지점을 찾기 위해 SKL 거리측정 함수를 이용하였다. SKL 거리측정 함수는 다음과 같이 정의되어진다.

$$D_{SKL}(P,Q) = \int (P(w) - Q(w))log(\frac{P(w)}{Q(w)})dw \quad (2)$$

where, $\int P(w)dw = 1$, $\int Q(w)dw = 1$

여기서 P(w)와 Q(w)는 <그림 3>에서 볼 수 있는 연속 합성단위들간의 연결 후보 영역에서의 정규화된 파워 스펙트럼이며 전체 면적은 1이다. P(w)와 Q(w)의 값이 같거나 유사하면 SKL 거리값은 0이나 0에 가까운 값을 가지며, 다를수록 큰 값을 가진다.

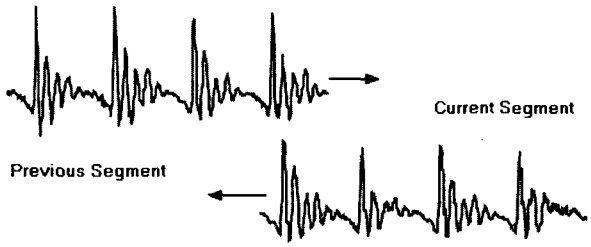
<그림 3>에서 짙은 회색의 직사각형이 음소 분할된 실제 합성단위의 영역이며 그 앞과 뒤에 옅은 회색의 직사각형은 여분의 음성 파형이며, 모든 합성단위는 이와 같이 일정한 길이의 여분의 정보를 포함하여 저장되어있다. <그림 3>에서와 같이 연결 예상 영역을 7개로 선정하고, 정규화된 파워 스펙트럼을 구한 뒤, 각 연결 쌍에서 가장 작은 스펙트럼 거리값을 가지는 영역을 선택하여 연결한다. 모든 구역이 아닌, 일부 후보 영역에 대해서만 거리측정을 하기 때문에 탐색 시간을 줄일 수 있다.

2.2.3. Phonetic Environment Checking

스펙트럼 불일치는 연결 경계에서 발견된다. 그러나 내부 실험 결과에 따르면, 연구개음이나 치경경구개음 중 하나로 분류된 음소가 모음과 연결될 때에는 불연속을 거의 인지 할 수 없다. 따라서 연결 합성단위의 좌우 음운환경을 고려하면 합성단위 연결시 신호처리를 생략할 수 있다. 이 경우 전체 수행속도가 짧아지는 장점이 있다. 이 음운환경에 따른 신호처리 생략에 관한 구체적인 내용은 차후 논문에서 소개할 계획이다.

2.2.4. Waveform Matching

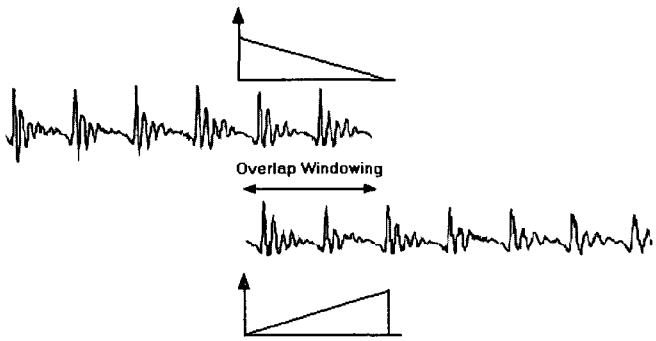
비록 합성단위의 경계에서 스펙트럼이 잘 일치하더라도, 동기화된 피치 연결을 보장하지는 못한다. 만약 양쪽 합성단위가 모두 유성음이라면, 음성 합성단위의 연결은 더욱 정확히 일치되어야 한다. 이미 스펙트럼의 일치 범위를 알고 있기 때문에, <그림 4>에서처럼 합성단위 슬라이딩 방법에 의해 피치 동기화된 일치점을 구할 수 있다. 그러나 무성음의 경우 대개 슬라이딩이 필요하지 않다. 합성단위의 슬라이딩은 하나의 피치 범위에 대해서, 상호상관(cross-correlation) 연산을 통하여 수행하였다.



<그림 4> 합성단위 슬라이딩을 이용한 최적의 정합지점 탐색

2.2.5. Overlap Windowing

만약 합성단위들의 레이블링이 정확하고, 연결된 합성단위들의 경계가 서로 정확히 일치한다면, 아무런 신호처리 없는 연결로도 높은 음질의 음성을 합성할 것이다. 그러나 일괄적인 자동 레이블링으로 합성단위들이 분할되어 있다면, <그림 5>에서와 같이 오버랩 창함수를 사용하여 더 나은 결과들을 얻을 수 있다. 사용된 창함수는 삼각창 함수이며, 시간영역에서 음성파형에 대해서 적용한 후 더함으로써 두 합성단위를 연결하였다. 이 때 창함수의 길이는 고정된 값을 사용하였으며, 약 두 개의 피치 주기를 사용하였다.



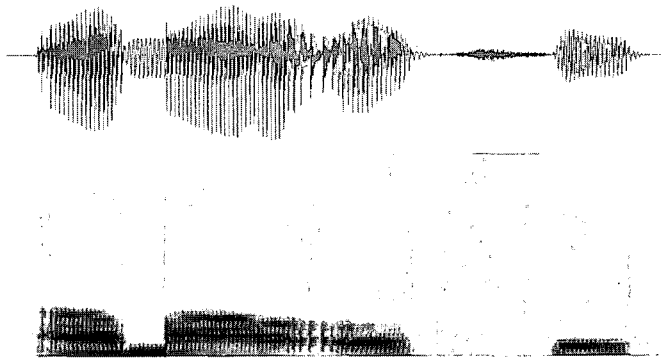
<그림 5> 삼각 창함수를 이용한 오버랩 합성단위 정합

3. 실험 및 결과

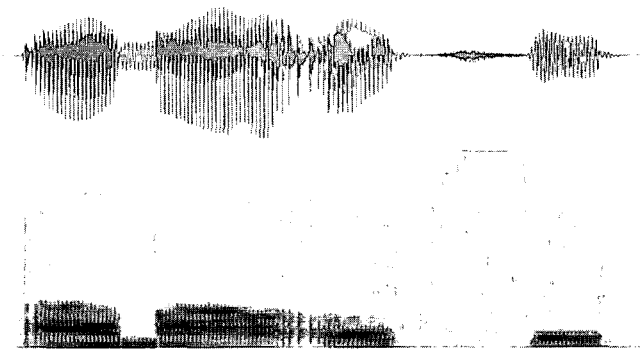
제안된 알고리즘의 유용성을 확인하기 위해 1부터 5 범위의 일반적인 MOS 테스트뿐만 아니라 ITU에서 권장하는 CCR(comparison category rating) 테스트도 사용하였다[8]. CCR 테스트에서는 서로 다른 두개의 합성 음성을 무작위로 청취자들에게 들려준다. 청취자들은 첫 번째 음성에 대해 두 번째의 음성의 음질이 얼마나 더 나은지 비교한 후 일곱 가지의 평가 중에서 하나를 선택한다. 이 일곱 가지 평가는 훨씬 더 낫다(3), 더 낫다(2), 조금 더 낫다(1), 거의 비슷하다(0), 조금 더 나쁘다(-1), 더 나쁘다(-2), 훨씬 더 나쁘다(-3)로 구성되어 있다. 본 실험에서는 더 낫다(1), 비슷하다(0), 더 나쁘다(-1)의 세 가지 평가를 사용하였다. 열명의 청취자들이 열개의 문장을 테스트하였으며, 열개의 문장 중 다섯 문장은 데이터베이스 구축시 사용된 문장이고, 나머지 다섯 문장은 신문에서 무작위로 선택된 문장이다. 실제 문장의 예는 <표 1>과 같다.

<표 1> 음성 합성 청취 테스트에 사용된 문장

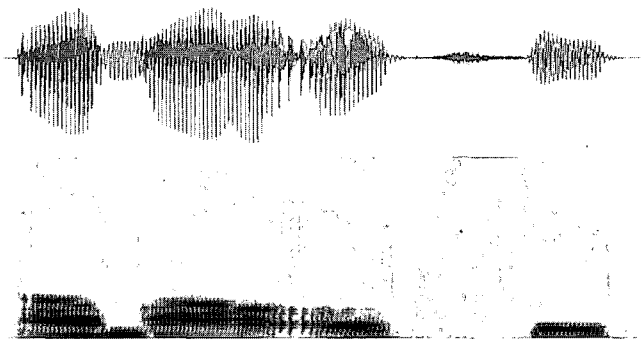
	실험에 사용된 문장
훈련 DB 내 문장	운전대를 잡은 사람치고 스치는 자동차를 한번쯤 욕하지 않은 사람은 아마 없을 것이다.
	요즘 애완용 개가 호강을 누리고 있는 것을 보면 앞으로 개팔자가 상팔자가 될 날도 정말 멀지 않았나 보다.
	이렇게 과보호 되다보니 요즘 애견들은 살만 피둥피둥 쪘지 건강실속은 예전 동네 골목을 쏘다니던 누렁이만도 못하다고 한다.
	새정부의 출범을 도왔고 개혁노선을 지지하는 장외의 개혁그룹들도 정치권의 물갈이를 요구하고있어 이들의 여권진입도 관측되고있다.
	올해의 금융 실명제 및 통화증발 휴유증 등이 겹쳐 당분간 물가상승을 억제하기가 힘들 것으로 전망했다.
훈련 DB 외 문장	이라크 아르빌에 주둔하고 있는 우리 자이툰 부대가 처음으로 로켓포 공격을 받았으나 인명피해는 없었습니다.
	일본 차기 총리로 유력한 아베 신조 일본 자민당 간사장 대리가 야스쿠니 신사참배는 총리의 당연한 책무라고 주장해 파문이 일고 있습니다.
	미국의 부시 대통령이 배아 줄기세포 연구 증진 법안에 대한 거부권 행사 의사를 다시 한번 명확히 했습니다.
	텍사스 레인저스의 박찬호가 팀의 8연승을 이끌며 시즌 5승과 함께 통산 99승을 달성했다.
	오늘도 대구가 29도까지 오르는 등 전국적으로 한여름 더위가 계속되겠습니다.



(a) Raw concatenation



(b) Overlap smoothing



(c) Proposed smoothing

<그림 6> 최종 합성음의 음성파형 및 스펙트럼: "아마 없을"

실험에 사용된 코퍼스 기반의 음성합성 시스템은 16KHz 샘플링 주파수와 16Bit 의 해상도를 가진 자동 음소 분할된 약 160,000개의 트라이폰(triphone) 음소들로 이루어진 음성데이터베이스를 사용한다.

서로 다른 세 종류의 연결 방법들을 비교하였으며, 실제 합성음의 음성파형 및 스펙트럼은 <그림 6>과 같으며, 청취 실험 결과는 <표 2>에 요약되어 있다. 여기서 overlap smoothing 방법은 앞 장에서 설명한 overlap windowing만을 적용한 것을 의미한다. <표 2>에서 두 가지 스펙트럼 평탄화 기술을 적용한 결과가 raw concatenation보다 좋다는 것을 확인할 수 있다. MOS 테스트 결과를 비교해 보면, 제안된 스펙트럼 평탄화 기술이 훈련된 데이터베이스에서나 신문 문장에서나 모두 raw concatenation이나 overlap smoothing 기술보다 성능이 뛰어남을 알 수 있다.

<표 2> MOS 및 CCR 테스트 결과

Smoothing Method	Test Sentence	MOS	CCR vote		
			Better	Similar	Worse
Raw Concatenation	Database	3.08	N/A	N/A	N/A
	Newspaper	2.78	N/A	N/A	N/A
Overlap Smoothing	Database	3.36	5	2	3
	Newspaper	3.14	5	3	2
Proposed Smoothing	Database	3.45	6	2	2
	Newspaper	3.27	5	3	2

4. 결 론

본 논문에서는 대용량 코퍼스 기반의 합성단위 연결 음성 합성 시스템을 위한 합성단위 연결 경계에서의 스펙트럼의 불연속성을 줄이는 스펙트럼의 평탄화 알고리즘을 제안하였다. 합성단위간 최적의 일치점을 찾기 위해 spectral matching과 waveform matching을 적용했고, 실제 합성단위들의 연결을 위해서 overlap windowing을 사용하였다. SKL 거리측정을 이용한 미리 정의된 후보 연결 쌍에 대한 spectral matching과 phonetic environment checking은 최적화된 연결점을 찾는 시간을 줄일 수 있다. 실험 결과를 통하여, 자동 음소분할로 구축된 음성 데이터베이스를 사용한 음성 합성 시스템에 제안된 알고리즘 적용시 적용하기 전보다 자연스러운 음성을 합성함을 확인할 수 있었다. 제안된 알고리즘은 임베디드 음성 합성기처럼 데이터베이스의 용량이 제한된 경우에도 유용할 것이라 생각된다.

참 고 문 헌

[1] A. J. Hunt, A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", *Proc. IEEE ICASSP*, pp.959-962, 1996.

[2] P. H. Low, S. Vaseghi, "Synthesis of unseen context and spectral and pitch contour smoothing in concatenated text to speech synthesis", *Proc. IEEE ICASSP*, pp.469-472, 2002.

[3] D. T. Chappell, J. H. L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis", *Speech Communication*, Vol. 36, Elsevier Science, pp.343-374, 2002.

[4] B. Pfister, "High-quality prosodic modification of speech signals", *Proc. ISCLP*, pp.2446-2449, 1996.

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

[6] E. Klabbers, R. Veldhuis, "On the reduction of concatenation artifacts in diphone synthesis", *Proc. ICSLP*, pp.1983-1986, 1998.

[7] E. Klabbers, R. Veldhuis, "Reducing audible spectral discontinuities", *IEEE Trans. on Speech and Audio Processing*, pp.39-51, 2001.

[8] X. Huang, A. Acero, H. Hon, *Spoken Lagnuage Processing*, Prentice Hall, pp.840-842, 2001.

접수일자 : 2005년 11월 30일
게재결정 : 2005년 12월 23일

▶ 김상진(Sang-Jin Kim)
주소: 305-714 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 공학부
전화: 042) 866-6206
E-mail: sangjin@icu.ac.kr

▶ 장경애(Kyung Ae, Jang)
주소: 137-792 서울시 서초구 우면동 17
소속: 한국통신(KT) 서비스개발연구소
전화: 02) 526-6755
E-mail: kajang@kt.co.kr

▶ 한민수(Minsoo Hahn) : 교신저자
주소: 305-714 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 공학부
전화: 042) 866-6123
E-mail: mshahn@icu.ac.kr