Spline 코드북 기반의 spectral folding을 이용한 대역폭 확장 방법

박 지 훈, 한 승 호, 양 희 식, 정 상 배, 한 민 수
한국정보통신대학교 음성/음향 정보연구실

Bandwidth Expansion Method Using Spline Codebook Based Spectral Folding

Jihoon Park, Seung Ho Han, Heesik Yang, Sangbae Jeong, Minsoo Hahn
Speech and Audio Information Lab., Information and Communications Univ.
E-mail : batho2n@icu.ac.kr

Abstract

Quality of narrowband speech (0~4kHz) can be enhanced by the bandwidth expansion technique, by which the high-band components are estimated. This paper proposes the bandwidth expansion method using the spline codebook based spectral folding. For the performance evaluation, the PESQ(Perceptual Evaluation of Speech Quality) scores are measured as the objective measurement. In addition, the MOS (Mean Opinion Score) and the preference tests are performed as the subjective measurement. The results show our proposed method outperforms the existing spline based one.

I. 서론

협대역 음성(0~4kHz)은 제거된 고주파 성분으로 인해 억눌린 음질과 명료도와 자연성이 떨어지는 음질 특성을 갖는다. 이러한 협대역 음성을 부호화하여 전송하는 음성 통신 시스템에서의 음질은 광대역 음성에 비해 크게 떨어진다. 대역폭 확장 방법은 협대역 음성에서 광대역 음성으로 대역폭을 확장 시키는 방법으로 협대역 음성을 분석하여 고주파 대역의 성분을 추정, 복원하는 방법이다. 대역폭 확장 방법을 음성 통신 시스템에 적용하여 협대역 음성을 광대역 음성으로 대역폭 확장을 수행한다면 향상된 음질을 얻을 수 있다.

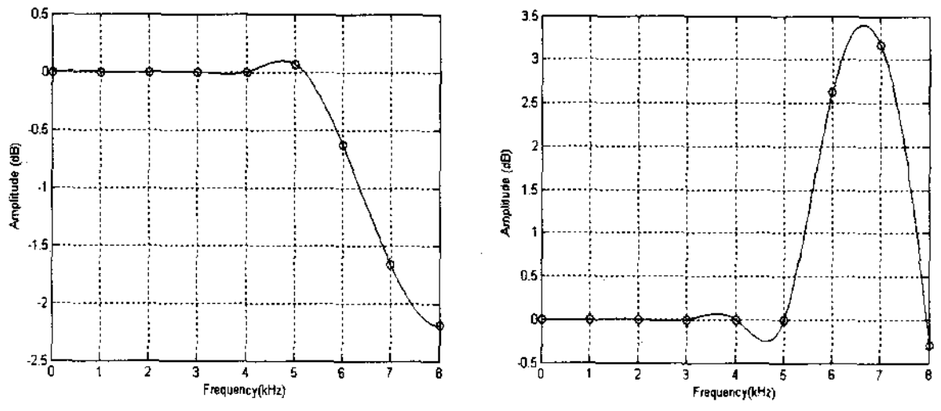
기존의 대역폭 확장 방법은 크게 스펙트럼 포락선 추정과 여기신호 생성 이라는 두 단계로 나누어진다. VQ(Vector Quantization)[1], GMM(Gaussian Mixture Model)[2], HMM(Hidden Markov Model)[3]은 스펙트럼 포락선을 추정하는 대표적인 방법들이다. 여기 신호를 생성하는 방법은 주기적 임펄스 입력과 혼합 여기신호[4] 등이 사용되고 있다. Spline을 이용한 spectral folding 방법[5]은 스펙트럼 포락선 추정과 여기신호 생성을 동시에 수행하여 대역폭을 확장하는 방법이다. 이 방법은 기존의 방법들에 비해 고주파 대역의 복원에 저주파 대역의 성분을 직접적으로 사용하기 때문에 복원되는 고주파 대역과 저주파 대역의 상관성이 크고, 실행시간도 짧아 실시간 실행에 용이하다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 cubic spline을 이용한 spectral folding 방법을 소개하고 3장에서는 제안하는 방법에 대해서 설명한다. 4장에서는 ITU-T 음성 코덱인 G.729A 로 부호화된 음성에 대해 대역폭 확장을 수행한 실험 결과를 객관적 평가방법인 PESQ와 주관적 평가 방법인 MOS, 선호도 테스트를 통해 보여주고 5장에서는 결론을 맺는다.

II. CUBIC SPLINE을 이용한 SPECTRAL FOLDING 방법

협대역 음성에 대해 Spectral folding을 적용하면 0~4kHz의 저주파 대역 정보가 4~8kHz의 고주파 대역에 4kHz를 기준으로 대칭하게 나타나면서 대역폭이

확장된다. Spectral folding의 문제점을 해결하고자 폴딩된 음성의 고주파 대역에 spline을 적용하여 음질을 개선하는 방법이 제안되었다[5]. Spline을 이용한 spectral folding 방법은 훈련 과정과 복원 과정으로 나눌 수 있다.



(a) (b)

그림 1. (a) 유성음일 경우의 spline 곡선
(b) 무성음일 경우의 spine 곡선

훈련 과정은 원음과 폴딩된 음성의 스펙트럼 포락선의 차이 값을 4, 5, 6, 7, 8 kHz의 다섯 점에서 구한다. 그리고 유성음, 무성음의 경우에 따라 각 주파수마다 차이 값의 평균을 구해 유성음, 무성음 각각의 대표 값들을 구한다. Cubic spline 방법을 사용해서 대표 값들의 사이 값을 추정하여 spline을 만든다. 만들어진 spline 곡선의 모양을 보면 그림 1과 같다.

복원 과정은 협대역 음성이 들어오게 되면 우선 spectral folding을 통해 음성의 대역폭을 확장하고 특징 추출을 통해 유성음, 무성음 분류를 수행한다. 푸리에 변환을 통해 주파수상으로 나타난 음성 신호에 훈련을 통해 만들어진 spline 곡선을 폴딩된 음성의 4000~8000Hz 고주파 대역 스펙트럼에 적용하여 고주파 대역의 성분을 복원한다.

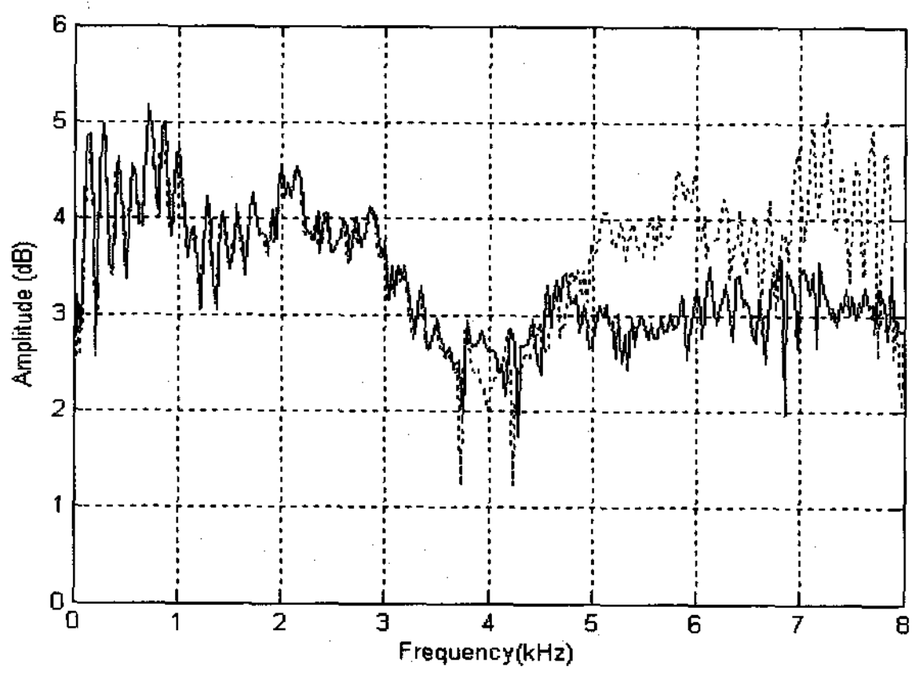


그림 2. /츠/ 음소의 스펙트럼 모양(실선: 원음성의 스펙트럼 모양, 점선: 폴딩된음성의 스펙트럼 모양)

그러나 유성음과 무성음 각 분류에서 spline 곡선 모양이 다양하게 나타남에 따라 고주파 대역을 완벽하게 복원하지 못하는 문제가 있다. 그림 2는 그 한 예로써 /츠/ 음소의 원음성 스펙트럼과(실선) 스펙트럼 폴딩된 음성의 스펙트럼 모양(점선)이다. 그림 2의 경우는 무성음이지만 폴딩된 음성의 고주파 대역 스펙트럼이 원음성의 고주파 대역 스펙트럼보다 증가된 유성음 경우의 특징을 보여준다. 그러나 음성 복원 과정에서는 무성음으로 분류되어 그림 1의 (b) spline이 적용된다. 이런 경우 복원된 고주파 성분이 원음의 스펙트럼과는 많은 차이를 보이게 되는 문제점이 발생한다.

III. 제안하는 대역폭 확장 방법

본 논문에서는 기존 대역폭 확장 방법의 문제점을 해결하기 위해 켑스트럼 벡터 양자화를 통해 spline 곡선 모양을 세분화하는 대역폭 확장 방법을 제안한다. 제안하고자 하는 시스템은 훈련 과정과 복원 과정으로 나눌 수 있다.

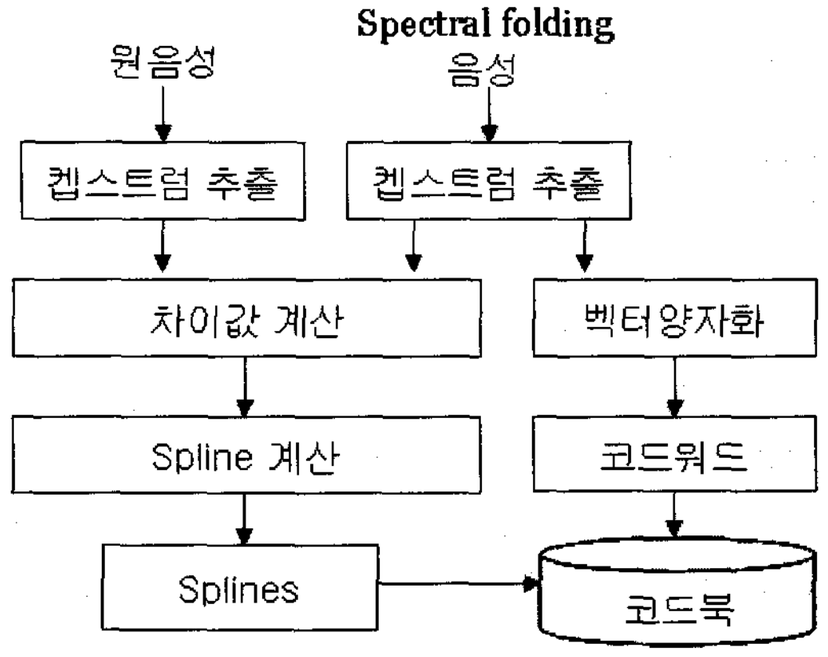


그림 3. 제안하는 방법의 훈련 과정 구성도

그림 3은 대역폭 확장방법의 훈련 과정 구성도이다. 원음성은 광대역 원음성이고 spectral folding 음성은 협대역 음성을 spectral folding한 음성이다. 각 음성의 분석구간에서 켑스트럼을 추출하고 폴딩된 음성의 켑스트럼 계수를 이용해서 벡터 양자화 훈련을 시켜 켑스트럼 코드북을 생성한다. 켑스트럼 계수를 이용해서 원음성과 폴딩된 음성의 스펙트럼 포락선을 구하게 되는데 켑스트럼 계수를 푸리에 변환 하게 되면 스펙트럼 포락선 정보를 알 수 있다. 원음성의 스펙트럼 포락선과 폴딩된 음성의 스펙트럼 포락선의 차이를 통해 spline 모양을 만드는데 필요한 대표 값을 구한다.

본 논문에서는 기존의 대역폭 확장 방법과 같이 4, 5, 6, 7, 8kHz 의 5점에서 대표 값을 사용하고 켑스트

럼 계수를 구하기 위해 FFT 켑스트럼을 사용한다. FFT 켑스트럼은 분석구간을 푸리에 변환하고 크기 스펙트럼을 구한다. 구해진 크기 스펙트럼에 로그 함수를 적용시키고 역 푸리에 변환을 하여 켑스트럼 계수를 구할 수 있다. spline 코드북은 켑스트럼 코드북의 코드워드(codeword) 별로 분류된 spline 대표 값들의 평균을 구한다. 구한 대표 값들을 cubic spline 방법을 사용해서 사이 값을 추정하여 spline 코드북을 생성한다.

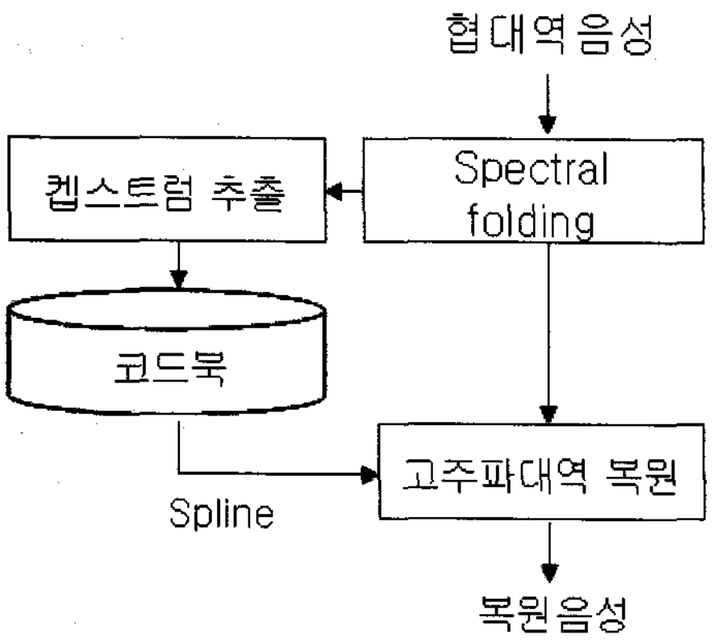


그림 4. 제안하는 방법의 복원 과정 구성도

복원 과정에서는 훈련 과정에서 만들어진 코드북을 사용해서 고주파 성분이 제거된 협대역 음성을 광대역 음성으로 복원한다. 그림 4는 대역폭 확장의 복원 과정 구성도이다. 협대역 음성이 들어오면 spectral folding을 취해주고 폴딩된 음성에서 켑스트럼 계수를 뽑아 코드북의 코드워드들과 상관성이 큰 코드워드와 대응되는 spline을 선택하게 된다. 선택된 spline을 푸리에 변환된 음성의 고주파 대역에 적용하여 제거되었던 고주파 대역의 성분을 복원한다. 데이터 기반의 분류를 통해 코드북을 생성하고 spline의 경우를 세분화하여 적용하는 splectral folding 방법을 제안함으로써 기존 방법의 문제점을 해결한다.

그림 2의 경우 기존 방법은 무성음의 경우임에도 폴딩된 음성의 고주파 대역 스펙트럼 모양이 원음성의 스펙트럼 모양보다 증가한 유성음 특징을 보여주는 결과가 나타났다. 그러나 제안한 방법으로 음성을 복원하게 되면 그림 5의 점선 모양과 같이 복원된 음성의 고주파 대역 스펙트럼 모양이 폴딩된 음성의 스펙트럼 모양보다 감쇄되는 것을 확인 할 수 있다. 이 결과는 제안된 방법이 기존 방법의 문제점을 해결하고 제안된 방법이 원음성의 스펙트럼과 유사성이 큰 것을 보여준다.

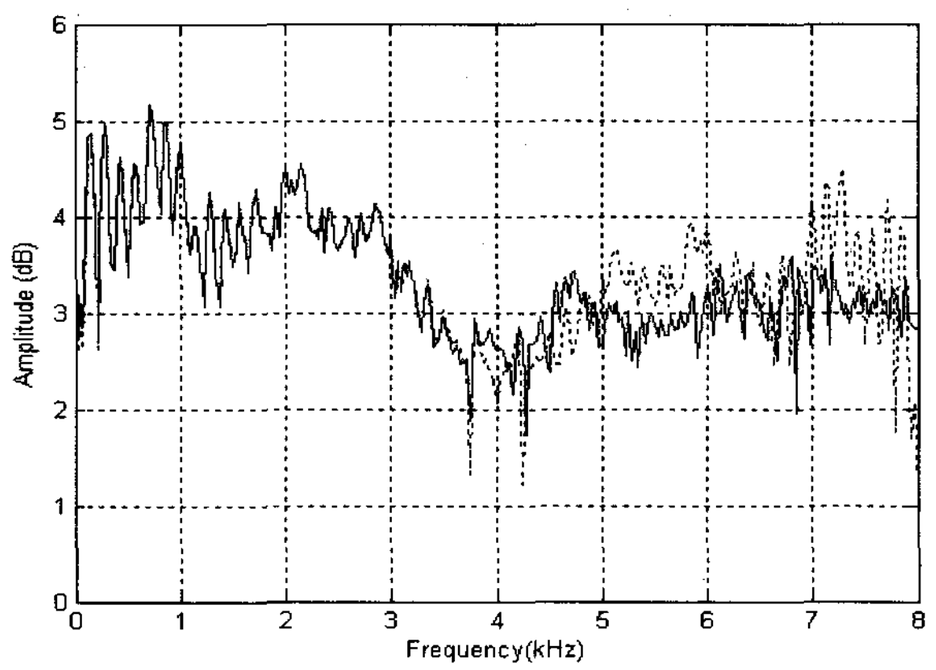


그림 5. /ㅊ/ 음소의 스펙트럼 모양(실선: 원음성의 스펙트럼, 점선: 제안한 방법으로 복원된 스펙트럼)

Ⅳ. 성능평가 및 결과

실험을 위해 10초 정도의 발성 길이를 가지는 한국어 낭독체 음성이 사용되었으며 여성 화자가 발성한 총 700문장의 데이터가 훈련과정에서 사용 되었다. 그리고 복원 과정에서는 훈련과정에서 사용된 문장이 아닌 다른 300문장이 사용되었다. 켑스트럼 벡터 양자화를 위해 LBG(Linde Buzo and Gray) 알고리즘[6]을 사용하였고 코드북 크기는 256으로 선택하였다. 복원된 음성의 음질 평가를 위한 객관적인 척도로서 PESQ[7] 테스트를 수행하고 주관적인 척도로서 MOS[8] 테스트와 음질 선호도 테스트를 수행하였다. 성능 평가 실험은 ITU-T 음성 코덱인 G.729A[9]로 부호화된 음성에 대해 분석구간 단위로 기존의 방법[5]과 제안된 방법으로 대역폭 확장을 실행하였다.

표 1. 기존의 방법과 제안된 방법의 PESQ 값

기준 음성	열화 음성	PESQ 값
원음성	기존 방법	2.53
원음성	제안된 방법	2.96

객관적 성능평가는 복원과정에서 사용된 300문장을 모두 사용하였고 음성 코덱 입력 신호인 광대역 음성을 기준 음성으로 정하였다. 코덱을 통과하면서 기존의 방법과 제안된 방법으로 대역폭이 확장된 음성을 열화 음성으로 정하여 PESQ 값을 산출하였다. 객관적 성능평가 결과는 표 1에서 보면 제안된 방법으로 복원된 음성의 PESQ 값이 기존의 방법으로 복원된 음성보다 0.43의 성능 향상 정도를 보인다.

주관적 성능 평가인 MOS 테스트와 선호도 테스트에는 10명의 청취자가 평가를 하였고 복원된 300문장 중에서 10문장을 무작위로 선택하여 기존의 방법으로

복원된 음성과 제안된 방법으로 복원된 음성의 MOS 값을 얻었다. MOS 테스트의 결과는 표 2와 같다. 표 2에서 보는바와 같이 기존의 방법으로 복원된 음성보다 제안된 방법으로 복원된 음성의 MOS 값이 0.5 정도의 성능 향상 정도를 보인다.

표 2. 기존의 방법과 제안된 방법의 MOS값

	MOS 값
기존방법	2.9
제안된 방법	3.4

표 3은 선호도 테스트 결과이다. 결과에서 보는바와 같이 복원된 음성의 음질을 선호하는 사람은 74%, 기존의 방법으로 복원한 음성의 음질을 선호한 사람은 22%이다. 선호도 없음은 두 가지 방법들로 복원된 음성 중 어떤 방법으로 복원된 음질이 나은지를 결정하지 못한 경우로써 4%였다. 다수의 사람들이 제안된 방법으로 복원된 음성을 선호한 것을 확인 하였다.

표 3. 기존의 방법과 제안된 방법의 선호도 결과

기존 방법	선호도 없음	제안된 방법
22%	4%	74%

객관적 성능평가 방법인 PESQ 값과 주관적인 성능평가 방법인 MOS, 선호도 테스트 결과를 보았을 때, 모두 기존의 방법으로 복원된 음성보다는 제안된 방법으로 복원된 음성의 음질이 향상됨을 확인하였다.

V. 결론

협대역 음성의 음질 향상을 위해 대역폭 확장 기술이 연구되어 왔다. 본 논문에서는 대역폭 확장 기술의 성능 향상을 위해 데이터 기반으로 켑스트럼 계수의 벡터 양자화를 이용해서 데이터 기반의 분류를 하고 고주파 대역의 성분을 복원하는 방법을 제안하였다. 그리고 PESQ 테스트, MOS 테스트와 선호도 테스트의 수행을 통해 음질이 향상되는 것을 확인하였다.

Spectral folding을 수행하면 저주파 대역의 하모닉 성분이 고주파 대역에 나타나게 되어 스펙트럼 포락선에는 변화를 주어도 하모닉 성분이 남아있게 되어 음질이 저하되는 문제점이 있다. 앞으로의 연구로써 고주파 대역의 하모닉 성분을 제거하는 방법에 대한 연구가 필요하다고 생각된다.

참고문헌

[1] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping", in *Proc. of ICSLP*, pp.1591-1594, 1994.

[2] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation", in *Proc. of ICASSP*, vol. 3, pp. 1843-1864, 2000.

[3] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden markov model", in *Proc. of IEEE Workshop on Speech Coding*, 2000.

[4] A. V. McCree and T.P.Barnwell III, "A mixed excitation LPC Vocoder Model for low bit rate speech coding", in *Proc. of IEEE Trans. On Speech and Audio Processing*, Volume: 3, Issue: 4, pp.242-250, July 1995.

[5] L. Laaksonen, J. Kontio and P. Alku, "Artificial bandwidth expansion method to improve intelligibility and quality of AMR-coded narrowband speech", in *Proc. of ICASSP*, pp.809-812, 2005.

[6] A. M. Kondoz, "Digital speech", John Wiley and Sons, 1994.

[7] ITU-T Recommendation, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders, *ITU-T Recommendation P.862*, Feb 2001.

[8] X. Huang, A. Acero and H-W. Hon, "Spoken language processing", Prentice Hall, 2001.

[9] ITU-T Recommendation, coding of speech at 8kbit/s using conjugate-structure algebraic-code -excited-linear-prediction (CS-ACELP), *ITU-T Recommendation G.729 Annex A*, Nov, 1996.