음성신호 압축 및 복원을 위한 음성 천이구간 검출과 근사합성 방식

이광석 · 김봉기 · 강성수 · 김현덕
진주산업대학교

Speech Transition Detection and approximate-synthesis Method for Speech Signal Compression and Recovery

Kwang-seok Lee · Bong-gi Kim · Seong-soo Kang · Hyun-deok Kim
Jinju National University
email : kslee@jinju.ac.kr

요 약

유·무성음의 음원을 이용한 음성부호화 시스템에서는 프레임 내에 유성자음과 무성자음이 공존하는 경우 음질의 왜곡을 수반할 수 있다. 따라서 프레임 내에 유성자음과 무성자음이 공존하지 않도록 하기 위해서 무성자음을 탐색 및 검출을 포함하는 천이구간을 제안하였다. 본 연구는 최소 자승법과 주파수 대역 분할법을 사용함으로써 TS 근사합성의 새로운 방식을 제시하였다. 결과적으로 이 방식은 0.547kHz이하와 2.813kHz 이상에서의 주파수 정보를 이용함으로써 TS내에서 고품질의 근사합성 파형을 얻을 수 있었다. 중요한 것은 최대 오류신호는 TS내에 저 왜곡 근사 합성파형이 생길 수 있다는 것이다. 이 방식은 유성음/묵음/TS의 새로운 음성부호화, 음성해석 및 음성합성에 적용할 수 있으리라 생각한다.

ABSTRACT

In a speech coding system using excitation source of voiced and unvoiced, it would be involved a distortion of speech quality in case coexist with a voiced and an unvoiced consonants in a frame. So, We proposed TS(Transition Segment) including unvoiced consonant searching and extraction method in order to uncoexistent with a voiced and unvoiced consonants in a frame. This research present a new method of TS approximate-synthesis by using Least Mean Square and frequency band division. As a result, this method obtain a high quality approximation-synthesis waveforms within TS by using frequency information of 0.547kHz below and 2.813kHz above. The important thing is that the maximum error signal can be made with low distortion approximation-synthesis waveform within TS. This method has the capability of being applied to a new speech coding of Voiced/Silence/TS, speech analysis and speech synthesis.

키워드

Speech Transition Detection, Approximate-Synthesis Method, Compression and Recovery

Ⅰ. 서 론

멀티미디어 콘텐츠 이용이 활발함에 따라 음성 또는 화상신호를 압축/복원하는 방식에 관심이 모아지고 있다. 음성신호를 압축/복원하는 음성부호화 방식에 있어서, 음성신호를 유성음(Voiced)/무성음(Unvoiced) 혹은 유성음 같은 선택정보에 의하여 유성음원과 무성음원을 유성음/무성음/무음(Silence)과 구동하여 음성신호를 재생하는 방식[1]~[5]에서는 음성신호를 수십ms의 고정된 프레임으로 분할하여 처리한다. 이때, 프레임 내, 음성신호가 유성음, 무성음, 무음과 같이 각기 독립적으로 존재하는 것이 아니라 무음(S)+무성음(UV) 또는 무음(S)+유성음(V), 유성음(V)+무성음(UV)의 형태로 존재하며, 이러한 형태의 음성신호는 과도기적인 특성을 나타낸다. 특히, 모음과 자음이 결합하여 유성음도 무성음도 아닌 특성을 나타내는 천이구간이 존재하는데, 이 천이구간의 음성신호를 유성음원이나 무성음원으로 재생하는 것은 문제점이라 볼 수 있다. 이러한 문제점을 해결하는 방법으로 유성음과 무성자음이 같은 프레임에 존재하지 않도록 프레임의 길이를 동적으로 할당하는 것도 고려해 볼 수 있으나, 이것은 디지털 신호처리의 특성상 상당히 어려운 처리 과정이라 할 수 있다.

본 연구에서는 특성을 달리하는 유성음부, 무음부, TS(Transition Segment) Including UnVoiced 자음부의 음성신호를 유성음(V), 무음(S), TS의 선택정보에 의하여 음성신호를 재생하는 V/S/TS 음성부호화 방식에 응용하기 위한 방법으로서, 제2절에서는 연속음성에서 V, S, TS를 탐색/추출한 다음, 프레임을 재구성함으로서 V/UV 방식에서의 음원 선택에 의한 문제점을 해결하는 방법을 제시한다. 제3절에서는 TS를 근

사 합성하는데 유효한 주파수 대역을 선택하여 근사 합성하는 방법을 제시하고, 제4절에서는 최소 자승법에 의하여 근사 합성하는 방법에 관하여 기술하고자 한다. 제5절에서 근사합성에 유효한 주파수 대역을 선택하는 방법과 최소 자승법에 의하여 근사 합성하는 방법의 비교검토 및 결론을 맺고자 한다.

II. V/S/TS 탐색 · 추출 및 분석

연속음성을 프레임으로 처리할 때 모음(V)과 무성자음(UVC)을 판정하기 위한 유효한 정보로 피치와 ZCR(Zero Crossing Rate)이 있는데, 이러한 정보를 상호 이용하면 V와 UVC, TS를 쉽게 판독할 수 있다. 피치정보를 추출하는 방법에는 프레임 단위로 정규화 된 피치정보를 추출하는 방법[6]-[8]이 있으나, 본 연구에서는 모음과 자음의 결합에 의하여 비교적 불규칙적으로 변화하는 TS의 위치를 효과적으로 탐색하고 추출하기 위하여 FIR필터와 STREAK필터를 혼합한 형태의 FIR-STREAK 디지털 필터로 음성신호를 처리하여 얻은 펄스성 잔차신호(R_P)로부터 개별 피치펄스의 위치를 추정한다.[1],[3],[4]

남여 10명 39문장의 연속음성을 관찰한 결과, V에서는 낮은 ZCR과 피치정보를 갖고 있고, UVC에서는 높은 ZCR과 피치정보가 없으며, 천이구간(TS)에서는 낮은 ZCR과 피치정보가 없는 특징을 나타내는 것을 알 수 있었다. 아울러, 연속음성에서 유성음의 지속시간은 100ms~500ms 정도이며 약2.7ms~12.5ms 간격마다 유사한 음성파형이 주기적으로 반복되는 특징을 갖고 있다. 반면, 무성자음의 경우는 무성 파열자음, 무성 마찰자음, 무성 파찰자음 별로 약간의 차이는 있으나 대개 20ms전후이고, 천이구간의 경우는 약 5ms전후인 지속시간을 갖는다.

이러한 특징들을 고려하여 연속음성에서 V, S, TS를 탐색하여 추출하는 방법을 그림1에 나타내었다. 이 방법에 있어서 음성신호는 3.4kHz LPF로 주파수 대역을 제한한 다음 10kHz-12bit로 표본화 및 양자화고, FFT 처리를 위하여 프레임의 길이는 25.6ms로 하였다.

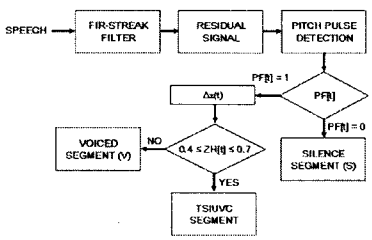

그림 1. TS의 탐색 및 검출
Fig.1 Search and Extraction of TS

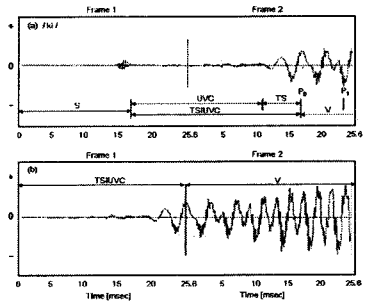

그림 2. V/S/TS 프레임 재구성
(a) 원 프레임 (b) 재구성한 프레임
Fig. 2 Frame Recomposition of V/S/TS
(a) Original frame (b) Reconstructed frame

그림1에서 프레임 안에 개별 피치정보가 하나라도 존재하지 않으면(PF(t)=0)) 프레임을 S로 판정하였고, 아니면 해당 프레임의 ZCR($Z[t]$)과 프레임간의 ZCR($\Delta Z[t]=Z[t]-Z[t-1]$)차, 천이구간(TS)과 무성자음구간(UVC)의 ZCR($ZH[t]$)이 $\Delta Z[t]<0, Z[t-1]\geq 0.4,\ 0.4\leq ZH[t]\leq 0.7$인 조건을 만족한 경우에 P_0위치에서 25.6ms 이전의 음성신호를 TS로 판정하였고, 그렇지 않다면 V로 판정하였다. 여기에서, 최초의 피치펄스(P_0)는 유성음의 시작위치인 동시에 TS가 끝나는 위치를 나타내는 중요한 정보이다. 결국, V, S, TS 판독한 결과를 근거로 그림2와 같이 프레임을 재구성함으로서 V/S/TS 처리에 적합한 신호처리 방법을 선택할 수 있도록 하였다.

III. 주파수 대역 선택에 의한 근사합성

제2절에서 탐색 · 추출한 TS를 재생하는데 유효한 주파수 대역을 선택하여 근사 합성법을 그림3에 나타내었다. 여기에서는 TS 재생에 유효한 주파수 대역을 알아보기 위하여 TS의 SNR및 스펙트럼 분석을 하였다.[4]

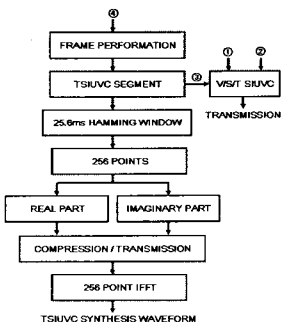

그림 3. 주파수 대역 선택에 의한 TS 근사합성법
Fig. 3 Approximate-Synthesis Method of TS using Frequency Band Selection

즉, TS 주파수 대역을 여러 개의 주파수 대역으로 분할하고, 각 주파수 대역의 신호를 사용하여 재생된 신호와의 SNR를 측정함으로서 근사합성에 유효한 주파수 대역을 선별하는 것이다.

우선, 10kHz로 표본화된 연속음성신호를 3.4kHz의 LPF로 주파수 대역을 제한하고, 프레임 당 256개의 샘플을 사용하였다. 또한 연속음성에서 추출한 TS를 스펙트럼 상에서 신호처리하기 위해 256-point 해밍창과 FFT를 사용하였으며, TS 스펙트럼을 29개의 대역으로 나누고, 각 대역의 주파수 정보를 IFFT하여 재생한 신호와의 SNR를 측정하였다. 여기에서 29개의 주파수 대역은 본 연구에서 사용하는 음성샘플을 FFT하면 5kHz 주파수대역에서 128개의 주파수 샘플이 존재하나, 3.4kHz의 주파수 대역제한을 하였기 때문에 실제로 87개의 주파수 신호를 사용하게 된다. 이때 각 주파수 신호의 간격이 Δf=39.0625Hz 가 되고, 최소 3개의 주파수를 사용하면 총 3.4kHz 주파수 대역은 29개의 주파수 대역으로 분할 할 수 있다. 여기에서, 사용하는 주파수의 개수는 제한된 것이 아니며, 경우에 따라서 1개 또는 2개의 주파수 신호를 사용할 수 있다. 다만, 본 연구에서 3개의 주파수를 사용하는 것은 제4장에서의 최소 자승법에서 근사곡선을 얻기 위한 적어도 3개의 데이터가 필요하였기에 데이터의 비교 분석을 위하여 사용하는 주파수의 수를 3개로 조정하였다.

SNR실험에 남여 9명의 대화체 음성(73문장, 무성 자음수: 195개)신호를 사용하였으며, 한 예로 무성자음의 SNR을 그림4에 나타냈다. 여기에서 주목할 것은 0.547kHz 이하의 낮은 주파수 대역과 2.813kHz 이상의 높은 주파수 대역에서 상대적으로 높은 SNR를 얻을 수 있었는데, 0.547kHz이하에서는 1.24~1.82dB이고, 2.813kHz 이상에서는 0.65~0.9dB를 얻을 수 있었다. 이것은 이번 실험을 통하여 유성음(V)의 주요 주파수 정보가 주로 400Hz이하의 낮은 주파수 대역에 분포하고, 무성자음(UVC)가 2~3kHz 부근의 높은 주파수 대역에 분포하고 있으며, 천이구간(TS)이 500Hz 부근의 중간 주파수 대역에 분포하고 있다는 것을 간접적으로 알 수 있었다. 특히, TS의 주요 주파수 정보가 높은 주파수와 중간 주파수대역으로 양분되어 있는 것을 알 수 있었다.

이러한 결과는 일반적으로 우리가 알고 있듯이 유성음(V)은 1kHz이하의 낮은 주파수 대역에 분포하고 있으며, 무성자음(UVC)은 2kHz이상의 높은 주파수 대역에 분포하고 있다는 사실과 일치하는 결과일 것이다. 다만, 천이구간(TS)의 주요 주파수 정보가 유성음(V)과 무성자음(UVC)의 중간 주파수 대역인 500Hz 부근에 존재한다는 사실에 주목할 필요가 있다.

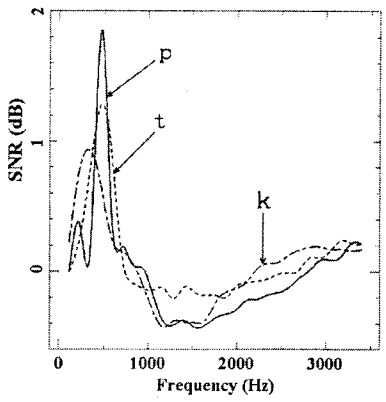

그림 4. TS 주파수대역의 SNR
Fig. 4 SNR of TS Frequency Band

Ⅳ. 최소자승법에 의한 근사합성

일반적으로 TS 신호는 유성음과 무성자음의 신호와 달리 신호의 진폭이 급격하게 변화하는 특성을 갖고 있기 때문에 선형적인 처리방법인 최소 자승법으로 TS를 처리할 경우에 많은 오차신호가 발생하게 되는데, 이러한 오차신호는 원래의 신호와 추정한 신호의 차로서 필요 없는 신호가 아니라 원래의 신호에 포함되어 있는 중요한 신호라고 볼 수 있다. 따라서 최소 자승법에 의하여 나타난 오차신호 중에 최대 오차신호 $e(x_{ij})$의 위치에 있는 주파수 신호(k)를 사용함으로서 TS 파형의 일그러짐을 보상할 수 있다고 생각되며, 실제로 사용하는 k의 수에 따라서 근사합성 파형의 보상정도가 달라지는 결과를 얻을 수 있었다. 여기에서 근사합성 파형의 일그러짐 보상 정도를 관찰할 수 있는 파형을 예로 들어 제시할 수 있으나, 파형의 일그러짐 보상 정도를 보다 통계적으로 평가할 수 있는 오차신호의 평가치를 사용하기로 하였다. 최소 자승법을 적용한 TS 근사합성법(Approximate-Synthesis Method)을 그림5에 제시하였으며, 이를 구체적으로 살펴보자. 우선, 신호의 진폭이 급격하게 변화하는 TS 신호의 특성을 고려하여 3.4kHz 주파수 대역에 존재하는 87개 주파수신호 x_{ij}를 다음과 같이 블록화 한다.

$$x_{ij} = (x_{00}, x_{01}, \ldots, x_{0M}) + (x_{10}, x_{11}, \ldots, x_{1M}) \\ \cdots\cdots\cdots\cdots \\ + (x_{m0}, x_{m1}, \ldots, x_{mM}) \quad (1)$$

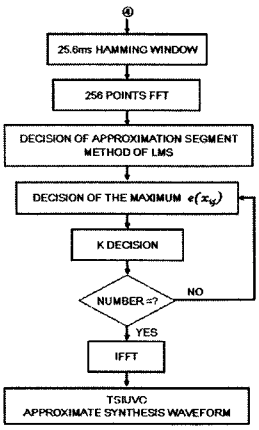

그림 5. MSE를 이용한 TS의 근사합성법
Fig. 5 Approximate-Synthesis Method of TS using MSE

이 블록화한 신호에 대한 근사신호 $y(x)$를 x의 n차 다항식으로 나타내면 다음과 같이 나타낼 수 있다.

$$y(x) = a_{00} + a_{01}x + a_{02}x^2 + \ldots + a_{0M}x^n$$
$$+ a_{10} + a_{11}x + a_{12}x^2 + \ldots + a_{1M}x^n$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$+ a_{m0} + a_{m1}x + a_{m2}x^2 + \ldots + a_{mM}x^n \quad (2)$$

이때, 각 신호에 있어서 실제 측정한 측정신호 (f_{ij})와 최소 자승법에 의하여 추정한 근사신호 ($y(x_{ij})$)의 차가 오차신호($e(x_{ij})$)가 된다.

$$e(x_{ij}) = y(x_{ij}) - f_{ij} \quad (3)$$

(2)식에서 계수 a_{ij}는 식(4)의 오차신호 평가값 (A)이 최소가 되도록 a_{ij}에 대하여 편미분하여 얻을 수 있다.

$$A = \sum_{i=0}^{m}\sum_{j=0}^{M} e^2(x_{ij}) \quad (4)$$

이때 만약 $e(x_{ij}) = 0$이라면 f_{ij}에 대하여 오차 없이 $y(x_{ij})$를 얻은 결과로서, $e(x_{ij})$에는 근사신호를 보정하기 위한 정보가 포함되어 있다고 볼 수 없으나, 만약, $e(x_{ij}) \neq 0$이라면 $e(x_{ij})$에는 근사신호를 보정할 수 있는 정보가 포함되어 있다고 볼 수 있다. $e(x_{ij})$를 이용하여 오차신호의 평가 값(A)을 제어할 수 있는지의 여부를 알아보기 위하여, 우선 근사 다항식의 차수와 근사신호의 수를 결정하여야 하는데, 전자의 경우는, 87개 주파수신호에 최소 자승법을 적용할 때 근사 다항식의 차수가 $n \geq 3$에서 거의 같은 근사치를 얻을 수 있었기 때문에 $n = 3$으로 하였고, 후자의 경우는 최소 2개에서 최대 29개로 하였을 경우에 분할 가능한 블록 수는 최소 3블록에서 최대 43 블록이 된다.

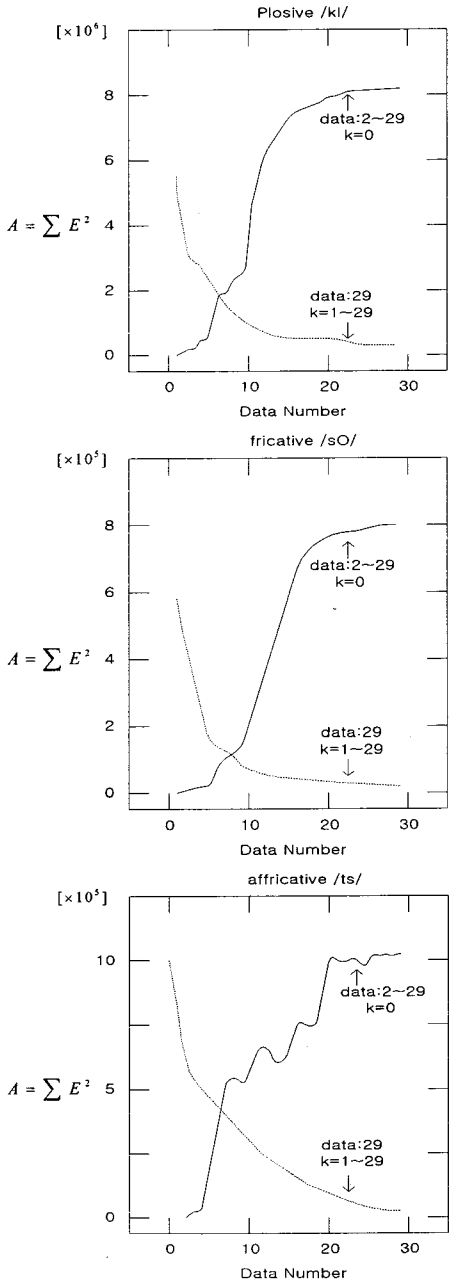

그림 6. MSE를 이용한 오차평가
Fig. 6 Error Performance using MSE

이때, 각 블록마다 측정한 오차신호($e(x_{ij})$)가 최대인 위치에 있는 주파수신호(k)를 사용하지 않은 경우(실선)와, 주파수신호(k)를 1~29개로 점차 늘리면서 적용한 경우(파선)의 오차신호 평가값(A)을 구하였다. 여기에서 사용한 음성 표본은 연속음성에서 자동 추출한 195개의 TS이며, 한 예로 그림6에 무성 자음의 오차신호 평가 값(A)을 나타냈다.

실험한 결과, k를 사용하지 않고 주파수 신호의 수를 증가시키면 오차신호의 평가 값(A)이 증가하는 것을 알 수 있는데, 이것은 데이터의 수가 커질수록 측정신호와 근사신호의 차가 커지는 것을 나타내는 것이다. 반면에 k를 점차 늘릴수록 오차신호의 평가 값(A)이 현저히 감소하는 것을 알 수 있다.

V. 결 론

과도기적인 특성을 나타내는 TS를 유성음원 또는 무성음원 어느 한쪽의 음원으로 재생하는 것은 무리가 있다고 생각된다. 그러므로 연속음성에서 TS를 탐색·추출한 다음, 유성음(V)/무음(S)/TS가 되도록 프레임을 재구성하고, TS를 근사 합성하는데 유효한 주파수 대역을 선택하는 근사합성법과 최소 자승법을 적용하여 오차신호가 최대인 위치에 있는 주파수신호(k)를 사용하는 근사합성법을 제안하였다.

실험 결과, 전자의 방법에서는 TS를 재생하는데 유효한 주파수 정보가 0.547kHz 이하와 2.813kHz 이상에 존재한다는 것을 알 수 있었으며, 후자의 방법에서는 오차신호가 최대인 위치에 있는 주파수신호(k)를 사용함으로서 원래의 파형에 근접한 근사합성 파형을 얻을 수 있었다. 이때, k의 수를 늘릴수록 파형의 일그러짐이 개선되었으며, 천이구간의 경우에는 $k \leq 5$, 무성자음의 경우에는 $k > 5$ 조건에서 양호한 근사합성 파형을 얻을 수 있었다. 제안한 방법들은 낮은 비트율의 유성음(V)/무음(S)/TS 선택의 음성부호화 방식에 적용하기 위한 전 단계이며 본 연구에서 제안한 방법들에 의한 음질 개선의 정도는 V/S/TS 음성부호화 방식을 구현하여 MOS 평가 등의 청각적 실험을 통하여 음질 개선의 정도를 정량적으로 측정하는 과제를 계속해 나갈 계획으로 있다.

참고문헌

[1]Hidffumi, Kobotake: "Optimization of Voiced/Unvoiced Decisions in Nonstationary Noice Envirionments", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-35, No.1, January 2000.

[3]武田昌一他:"殘差音源利用分析合成方式とマルチパルス法の基本特性の比較檢討", 電子情報通信學會論文誌, Vol.J73-A, No.11, 2004.

[4]眞野 淳,小澤 愼治: "LPC有聲音殘差のピッチ同期メルLSP分析合成方式", 電子情報通信學會論文誌, Vol.J71-A, No.3, 2005.

[5]武田昌一他:"殘差音源利用分析合成方式とマルチパルス法の基本特性の比較檢討", 電子情報通學會論文誌, Vol.J73-A, No.11, 2003.

[6]藤井健作:"自己相關法による電話帶域音聲のピッチ抽出法", 電子情報通信學會技術報告書, sp87-65, 2002.

[7]Chong Kwan Un,Shin-Chien Yang: "A Pitch Extraction Algorithm Based on LPC Inverse Filtering and AMDF", IEEE, Vol. ASSP-39, Feb, 2005.

[8]CarolA.McGonegal,Lawrence R.Rabiner,Aaron E.Rosenberg:"Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech", IEEE, Vol.ASSP-25, June, 2006.