주파수 분할 및 최소 자승법을 이용한 TSIUVC 근사합성법에 관한 연구

이 시 우†

요 약

유성음원과 무성음원을 사용하는 음성부호화 방식에 있어서, 같은 프레임 안에 모음과 무성자음이 있는 경우에 음질저하 현상이 나타난다. 본 연구에서는 같은 프레임안에 유성음과 무성자음이 존재하지 않도록 FIR-STREAK 필터와 zerocrossing rate을 이용한 개별피치 펄스를 사용하여 연속음성에서 무성자음을 포함한 천이구간(TSIUVC)을 탐색, 추출하는 방법을 제안한다. 또한 본 논문에서는 최송 자승법과 주파수 대역 분할을 이용한 TSIUVC 근사합성법을 제안하였다. 실험결과, 0.547kHz 이하 2.813kHz 이상의 주파수 정보를 사용하여 TSIUVC 음성파형을 양호하게 근사합성할 수 있었으며, 최대 오차신호가 일그러짐이 적은 TSIUVC 근사합성 파형에 중요한 역할을 한다는 것을 알 수 있었다. 이 방법은 음성합성, 음성분석, 새로운 Voiced/Silence/TSIUVC의 음성부호화 방식에 활용할 수 있을 것으로 기대된다.

A Study on TSIUVC Approximate-Synthesis Method using Least Mean Square and Frequency Division

See-Woo.LEE†

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, I propose TSIUVC(Transition Segment Including UnVoiced Consonant) searching and extraction method in order to uncoexistent with a voiced and unvoiced consonants in a frame. This paper present a new method of TSIUVC approximate-synthesis by using Least Mean Square and frequency band division. As a result, this method obtain a high quality approximation-synthesis waveforms within TSIUVC 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 TSIUVC. This method has the capability of being applied to a new speech coding of Voiced/Silence/TSIUVC, speech analysis and speech synthesis.

Key words: TSIUVC

1. 서 론

근래, 인터넷과 이동통신의 이용자가 급증함에 따라 음성 또는 화상신호를 압축/복원하는 방식에 관심이 모아지고 있다.

본 논문은 상명대학교 2001학년도 계당장학회 연구비 지원 / 2002학년도 교내 연구비 지원에 의하여 연구되었음.
접수일 : 2002년 12월 9일, 완료일 : 2003년 1월 14일
† 정회원, 상명대학교 정보통신전공 교수

음성신호를 압축/복원하는 음성부호화 방식에 있어서, 음성신호를 유성음(Voiced)/무성음(Unvoiced) 혹은 유성음(Voiced)/무성음(Unvoiced)/무음(Silence)과 같은 선택정보에 의하여 유성음원과 무성음원을 구동하여 음성신호를 재생하는 방식[1-4]에서는 음성신호를 수십ms의 고정된 프레임으로 분할하여 처리한다. 이때, 프레임내 음성신호가 유성음, 무성음, 무음과 같이 각기 독립적으로 존재하는 것이 아니라 무음(S)+무성음(UV) 또는 무음(S)+유성음(V), 유성음(V)+무

성음(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 처리과정이라 할 수 있다.

본 논문에서는 특성을 달리하는 유성음(Voiced)부, 무음(Silence)부, TSIUVC(Transition Segment Including UnVoiced Consonant)부의 음성신호를 유성음(V), 무음(S), TSIUVC의 선택정보에 의하여 음성신호를 재생하는 V/S/TSIUVC 음성부호화 방식에 응용하기 위한 방법으로서, 제2장에서는 연속음성에서 V, S, TSIUVC를 탐색/추출한 다음, 프레임을 재구성함으로서 V/UV 방식에서의 음원 선택에 의한 문제점을 해결하는 방법을 제시한다. 제3장에서는 TSIUVC를 근사합성하는데 유효한 주파수 대역을 선택하여 근사합성하는 방법과 최소 자승법에 의하여 근사합성하는 방법에 관하여 기술하고자 한다. 제 4장에서 본 논문에서 제시한 방법들의 실험결과에 관하여 기술하고, 제5장에서 결론을 맺고자 한다.

2. V/S/TSIUVC 탐색 · 추출 및 분석

연속음성을 프레임으로 처리할 때 모음(V)과 무성자음(UVC)을 판정하기 위한 유효한 정보로 피치(pitch)와 ZCR(zero crossing rate)이 있는데, 이러한 정보를 상호 이용하면 V와 UVC, TSIUVC를 쉽게 판독할 수 있다.

피치정보를 추출하는 방법에는 프레임 단위로 정규화된 피치정보를 추출하는 방법[5-7]이 있으나, 본 연구에서는 모음과 자음의 결합에 의하여 비교적 불규칙적으로 변화하는 TSIUVC의 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.[8]

남여 9명 73문장의 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~500ms정도이며 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 20ms 전후이고, 천이구간의 경우는 약 5ms 전후인 지속시간을 갖는다.

이러한 특징들을 고려하여 연속음성에서 V, S, TSIUVC를 탐색하여 추출하는 방법을 그림 1에 나타냈다.

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 1을 간략히 설명하면, 프레임 안에 개별 피치정보가 하나라도 존재하지 않으면(PF[t]=0) 프레임을 S로 판정하였고, 그렇지 않다면 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 TSIUVC로 판정하였고, 그렇지 않다면 V로 판정하였다. 여기에서 제시한 임계값들은 남여 9명 73문장의 음성샘플(모음:609개, 무성자음:195개)에서 추출한 TSIUVC 샘플에 대한 ZCR을 관찰한 결과에 의하여 얻어진 임계값이다. 이와 같은 임계값의 조건과 본래의 음성샘플에 TSIUVC가 존재함에도 불구하고 추출되지 않았을 경우(b)와 TSIUVC가 존재하지 않는데도 불구하고 추출된 경우(c)를 TSIUVC추출

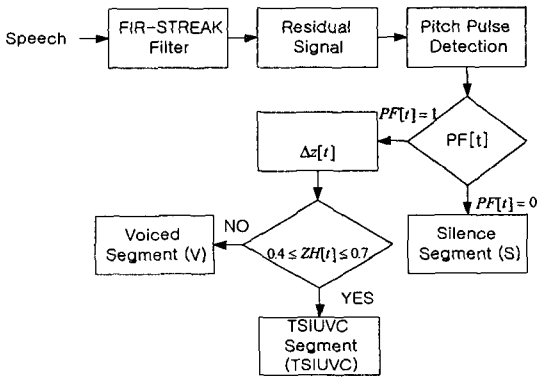

그림. 1 TSIUVC의 탐색과 추출법

오류로 규정한 (1)식에 의하여 TSIUVC 추출율을 산출하였다.

$$\Re = \sum_{j=1}^{m} \{a_j - (b_j + c_j)\} / \sum_{j=1}^{m} a_j \qquad (1)$$

aj : TSIUVC 관찰수, m : 음성샘플 수

실험결과, 남자음성에서 96.2%, 여자음성에서는 91%의 TSIUVC 추출율을 얻을 수 있었다. 단, 이와 같은 추출율은 상기의 ZCR 임계값에 따라서 추출율이 달라질 수는 있다.

여기에서, 최초의 피치펄스(P_0)는 유성음의 시작 위치인 동시에 TSIUVC가 끝나는 위치를 나타내는 중요한 정보이다. 결국, V, S, TSIUVC 판독한 결과를 근거로 그림 2와 같이 프레임을 재구성함으로서 V/S/TSIUVC 처리에 적합한 신호처리 방법을 선택할 수 있도록 하였다.

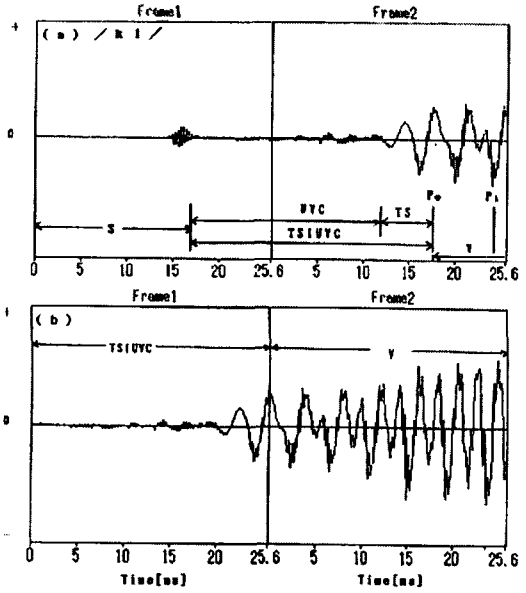

그림. 2 V/S/TSIUVC의 프레임의 재구성
(a) 본래의 프레임 (b) 재구성한 프레임

3. TSIUVC 근사합성

3.1 주파수대역 선택

제 2장에서 탐색·추출한 TSIUVC를 재생하는데 유효한 주파수 대역을 선택하여 근사합성하는 방법(Approximate-Synthesis Method)을 그림 3에 나타내었다. 여기에서는 TSIUVC 재생에 유효한 주파수 대역을 알아보기 위하여 TSIUVC의 SNR및 스펙트럼 분석을 하였다. 즉, TSIUVC 주파수 대역을 여러개의 주파수 대역으로 분할하고, 각 주파수 대역의 신호를 사용하여 재생된 신호와의 SNR를 측정함으로서 근사합성에 유효한 주파수 대역을 선별하는 것이다.

10kHz로 표본화된 연속음성신호에서 추출한 TSIUVC를 스펙트럼 상에서 신호처리하기 위해 256 point Hamming Window와 FFT를 사용하였으며, TSIUVC 스펙트럼을 29개의 대역으로 나누고, 각 대역의 주파수 정보를 IFFT하여 재생한 신호와의 SNR를 측정하였다.

여기에서 주파수 대역을 29개로 선정한 이유는, 본 연구를 통신 시스템에 적용하였을 경우를 고려하여 3.4kHz의 LPF로 주파수 대역을 제한하였으며, 음성신호를 10kHz로 표본화하였기 때문에 주파수 간격이 Δf=39.0625Hz 이 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 최소 자승법에서 근사곡선을 얻기 위한 최소 데이터수가 3개이기 때문에 이 방법들의 비교검토가 용이하도록, 사용하는 주파수의 수를 조정하였다.

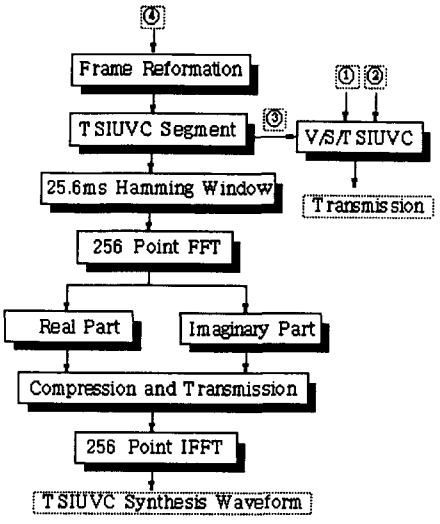

그림. 3 주파수 대역 선택에 의한 TSIUVC 근사합성법

3.2 최소 자승법

최소 자승법을 적용한 TSIUVC 근사합성법(Approximate-Synthesis Method)을 그림 4에 제시하였다. 일반적으로 TSIUVC 신호는 유성음과 무성자음의 신호와 달리 신호의 진폭이 급격하게 변화하는 특성을 갖고 있기 때문에 선형적인 처리방법인 최소 자승법으로 TSIUVC를 처리할 경우에 많은 오차신호가 발생하게 되는데, 이 오차신호 중에 최대 오차신호 $e(x_{ij})$의 위치에 있는 주파수 신호(k)를 사용하여 TSIUVC 파형의 일그러짐을 보상하게 되며, 이때 사용하는 k의 수에 따라서 근사합성 파형의 보상정도가 달라지게 된다.

우선, 신호의 진폭이 급격하게 변화하는 TSIUVC 신호의 특성을 고려하여 TSIUVC의 주파수 신호를 블록화한 후 최소 자승법을 적용하고자 하였다.

TSIUVC 신호를 FFT하면 3.4kHz 주파수 대역에 87개 주파수신호가 존재하게 되며, 이 주파수 신호 x_{ij}를 다음 식과 같이 블록화 한다.

$$\begin{aligned} x_{ij} = &(x_{00}, x_{01}, \ldots, x_{0M}) \\ &+ (x_{10}, x_{11}, \ldots, x_{1M}) \\ &\cdots\cdots\cdots\cdots \\ &+ (x_{m0}, x_{m1}, \ldots, x_{mM}) \end{aligned} \quad (2)$$

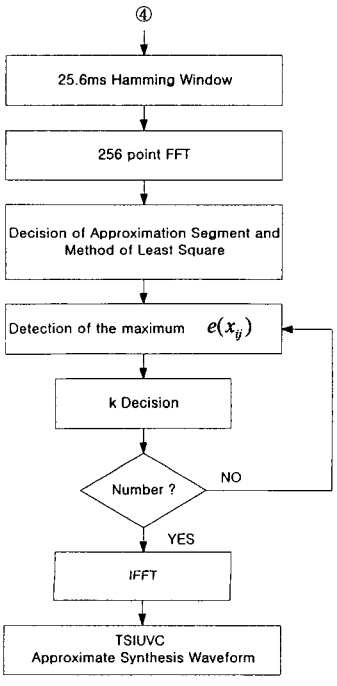

그림. 4 최소 자승법에 의한 TSIUVC 근사합성법

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

$$\begin{aligned} y(x) = &\ a_{00} + a_{01}x + a_{02}x^2 + \ldots + a_{0M}x^n \\ &+ a_{10} + a_{11}x + a_{12}x^2 + \ldots + a_{1M}x^n \\ &\cdots\cdots\cdots\cdots \\ &+ a_{m0} + a_{m1}x + a_{m2}x^2 + \ldots + a_{mM}x^n \end{aligned} \quad (3)$$

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

$$e(x_{ij}) = y(x_{ij}) - f_{ij} \quad (4)$$

그리고, (3)식에서 계수 a_{ij}는 식(5)의 오차신호 평가값(A)이 최소가 되도록 a_{ij}에 대하여 편 미분하여 얻을 수 있다.

$$A = \sum_{i=0}^{m} \sum_{j=0}^{M} e^2(x_{ij}) \quad (5)$$

이때 만약 $e(x_{ij})$=0이라면, 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. 결과적으로, 이러한 정보를 이용하여 TSIUVC 근사합성 파형의 일그러짐을 보상하게 되는데, 일그러짐의 보상정도는 상기 (5)식의 평가 값을 측정함으로서 알 수 있다. 이 평가값에 관해서는 제4장에서 언급하고자 한다.

4. 실험결과

주파수 선택과 최소 자승법은 TSIUVC를 근사합성하는데 사용하는 파라미터의 정보량이 다르기 때문에 품질적 측면의 비교평가 보다는 TSIUVC를 근사합성하는데 유효한 정보가 무엇인지를 탐구하는 관점에서 이루어졌다.

남여 9명의 대화체 음성(73문장, 무성자음수:195개) 샘플을 사용하여 추출한 TSIUVC를 주파수 선택 및 최소 자승법을 이용하여 SNR, 오차신호의 평가값 및 근사합성하는 실험을 하였다.

주파수 선택에 의한 방법은, 음성샘플에서 자동 추출한 TSIUVC를 FFT하여 얻은 주파수 스펙트럼을 여러 대역으로 분할하고, 이들 대역을 각기 사용한 경우의 SNR을 분석하였다. 무성파열, 무성마찰 또는 무성

파찰 자음을 갖는 TSIUVC의 경우에 모두 0.547kHz 이하의 낮은 주파수 대역과 2.813kHz 이상의 높은 주파수 대역에서 상대적으로 높은 SNR를 얻을 수 있었다. SNR결과의 한 예를 그림 5에 나타냈다. 여기에서 다른 주파수 대역에 비하여 0.547kHz 이하와 2.813kHz 이상의 주파수 대역에서 상대적으로 높은 SNR을 나타내었는데, 이것은 TSIUVC의 주요 주파수 정보가 0.547kHz 이하와 2.813kHz 이상의 주파수 대역으로 양분되어 있음을 나타내는 것이다.

또한 이것은 일반적으로 유성음(V)과 무성자음(UVC)의 주요 주파수 정보가 주로 1kHz이하와 2kHz 이상의 주파수 대역에 분포하고 있다는 사실을 뒷받침하는 결과라 할 수 있다. 여기에서, 주목할만한 사실은 천이구간(TS)의 주요 주파수 정보가 500Hz 부근의 중간 주파수 대역에 분포하고 있다는 것이다. 이러한 사실은 그림 6에서도 확인 할 수 있었는데, 그림 6의 (a)~(f)는 0.547kHz 주파수 대역을 사용하여 근사합성한 파형이고, (g)~(k)는 (a)~(f)를 조합한 경우의 근사합성 파형이다. 여기에서 (a)의 경우에는 TS구간의 파형이, (f)의 경우에는 UVC구간의 파형이 잘 근사합성되어 있는 것을 알 수 있었다. 결국, (a)+(f)를 조합한 (k)의 경우에 본래의 파형에 근접한 TSIUVC 파형을 얻을 수 있었다.(여기에서 TSIUVC 본래의 파형은 (그림 8)의 (a)이다.)

최소 자승법은 (4)식의 $e(x_{ij})$를 이용하여 오차신호의 평가 값(A)을 제어할 수 있는지의 여부를 알아보기 위하여, 우선 근사 다항식의 차수와 근사신호의 수를

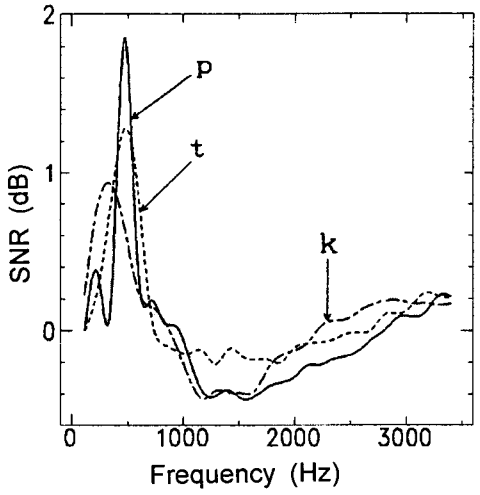

그림. 5 TSIUVC 주파수 대역의 SNR

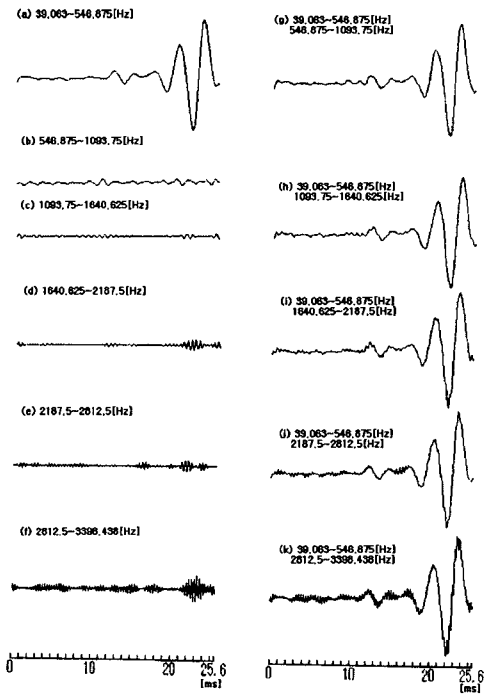

그림. 6 주파수 선택에 의한 TSIUVC 근사합성파형

결정하여야 한다. 전자의 경우는, 87개 주파수신호에 최소 자승법을 적용할 때 근사 다항식의 차수가 $n \geq 3$에서 거의 같은 근사값을 얻을 수 있었기 때문에 $n=3$으로 하였고, 후자의 경우는 근사신호를 얻기 위해서 최소한 2개 이상의 주파수 신호가 필요하기 때문에 근사신호의 수를 최소 2개로 하였다. 이때 근사신호의 수를 최대 29개로 하였을 경우에 분할 가능한 블록 수는 최소 3블록에서 최대 43블록이 된다. 이 블록을 점차 늘리면서 오차신호($e(x_{ij})$)가 최대인 위치에 있는 주파수신호(k)는 사용하지 않은 경우(실선)와, k를 1~29개 적용한 경우(파선)의 오차신호의 평가 값(A)을 구하였다. 그림 7에 오차신호의 평가 값(A)의 예를 나타내었는데, k를 사용하지 않고 주파수 신호의 수를 증가시킴에 따라서 오차신호의 평가 값이 증가하는 것을 알 수 있었다.

반면에 k를 점차 늘릴수록 오차신호의 평가 값이 현저히 감소하는 것을 알 수 있었다.

결국, k를 늘릴수록 오차신호의 평가 값이 점차 감소한다는 것은 근사합성 파형의 일그러짐이 점차 개선된다는 것을 나타내는 것이다. 이를 입증하는 예로서 그림 8에 k를 사용하지 않은 경우(k=0)와 k를 1~10개 사용한 경우를 나타내었다. 실험결과, k를 늘릴수록 파

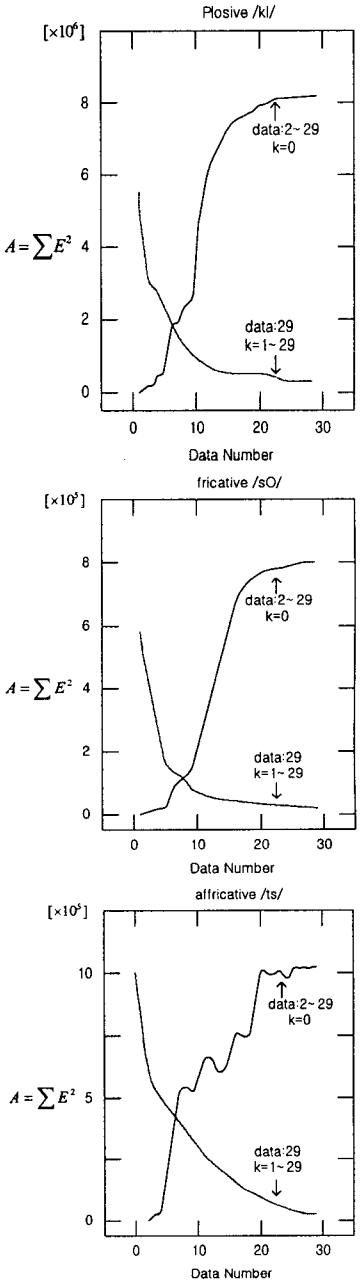

그림. 7 최소 자승법의 오차 평가

형의 일그러짐이 점차 개선되었는데, 그 개선의 정도가 천이구간의 경우에는 $k\leq5$, 무성자음의 경우에는 $k>5$에서 효과적임을 알 수 있었다.

5. 결 론

과도기적인 음성신호의 특성을 갖고 있는 TSIUVC

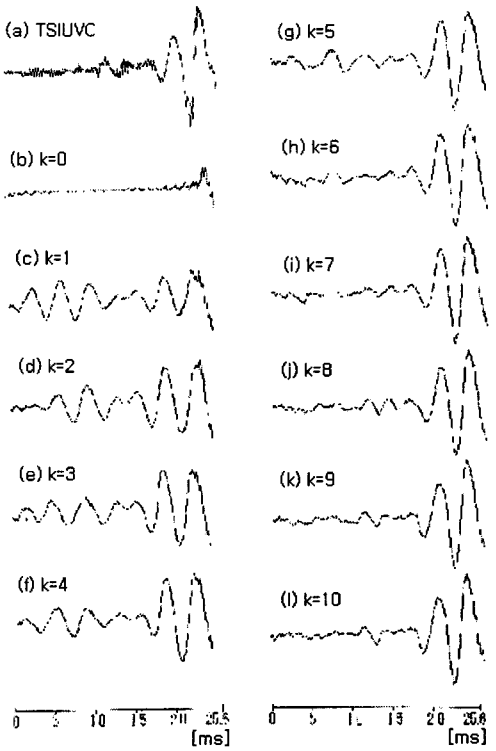

그림. 8 최소 자승법에 의한 TSIUVC 근사합성파형

를 유성음원 또는 무성음원 어느 한쪽의 음원으로 재생하는 것은 무리가 있다. 따라서, 본 연구에서는 연속음성에서 TSIUVC를 탐색·추출한 다음 프레임 안에 음성신호가 유성음(V)/무음(S)/TSIUVC가 되도록 프레임을 재구성한 다음, TSIUVC를 근사합성하는데 유효한 주파수 대역을 선택하는 근사합성법과 최소 자승법을 적용하여 오차신호가 최대인 위치에 있는 주파수신호(k)를 사용하는 근사합성법을 제안하였다.

실험결과, 전자의 방법에서는 TSIUVC를 재생하는데 유효한 주파수 정보가 0.547kHz 이하와 2.813kHz 이상에 있다는 것을 알 수 있었다. 후자의 방법에서는 오차신호가 최대인 위치에 있는 주파수신호(k)를 사용함으로서 원래의 파형에 근접한 근사합성 파형을 얻을 수 있었는데, 이때, k의 수를 늘릴수록 파형의 일그러짐이 점차 개선되었는데, 특히 천이구간의 경우에는 $k\leq5$, 무성자음의 경우에는 $k>5$ 조건에서 일그러짐이 개선되는 것을 알 수 있었다.

이번에 제안한 방법들은 낮은 bit rate의 유성음(V)/무음(S)/TSIUVC 선택의 음성부호화 방식에 적용하기 위한 전 단계로서 연구된 것이며, 본 논문에서 제안

한 방법들에 의한 MOS 평가는, 향후 V/S/ TSIUVC 음성부호화 방식을 구현하여 평가하기로 한다.

참 고 문 헌

[1] CHONG KWAN UN and HYEONG HO LEE: "Voiced/Unvoiced/Silence Discrimination of Speech by Delta Modulation", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No.4, August 1980.

[2] HIDEFUMI KOBATAKE: "Optimization of Voiced/Unvoiced Decisions in Nonstationary Noice Envirionments", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-35, No.1, January 1987.

[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, 1990.

[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, 1988.

[5] 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, sp 87-65. 1987.

[6] Chong Kwan Un and Shin-Chien Yang: "A Pitch Extraction Algorithm Based on LPC Inverse Filtering and AMDF" IEEE, Vol. ASSP-39, Feb, 1991.

[7] Carol A.McGonegal, Lawrence R.Rabiner and Aaron E.Rosenberg: "Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech", IEEE. Vol. ASSP-25, June, 1997.

[8] 이시우: "FIR-STREAK 디지털 필터를 사용한 피치추출 방법에 관한 연구", 한국정보처리, 학회논문지, 제6권 제1호, p247-252, 1999.1.

이 시 우

1987년 동국대학교 전자공학과 학사

1990년 日本大學(Nihon Univ) 전자공학과 석사

1994년 日本大學(Nihon Univ) 전자공학과 박사

1994년~1998년 삼성전자 통신연구소/멀티미디어 연구소

1998년~현재 상명대학교 정보통신전공 교수

관심분야 : 음성신호처리, 유무선통신

교 신 저 자

이 시 우 330-720 충남 천안시 안서동 98-20 상명대학교 정보통신전공 교수