PSOLA 알고리즘을 이용한 친절전화기능의 구현에 관한 연구

정 현 욱, 김 종 국, 배 명 진
숭실대학교 정보통신공학과

A Study on a Implementation of Gentle Phone`s Fuction by using PSOLA Algorithm

HyunUk Jung, JongKuk Kim, MyungJin Bae
Dep. of Information and Telecom. Engr. Soongsil Univ.
t2studio@nate.com

요 약

본 논문은 전화기의 수화기에서 들리는 상대방의 목소리를 디지털 발성처리기술을 적용하여 억양이 강하지 않고 부드러운 소리(소프트사운드, soft-sound)로 통화하는 방식을 새로이 제안한 것이다. 실시간 친절전화기의 구현에 있어 메모리 점유율을 음성신호의 지속시간을 제어함으로써 효율적인 소프트웨어 및 하드웨어 구현을 위한 방법을 제안한다. 목소리 신호의 특징 추출을 수행하여 발성자의 특성정보는 그대로 유지하면서 발성자의 의미정보를 친절하게 변경하는 것으로서, 발성자의 발성특성에서 지속시간을 조절하여 슬로우-목소리를 구현하거나, 발성 지속시간의 지연을 유성 및 비유성 구간으로 구분하여 처리를 다르게 하는 등의 발성변환법을 전화기에 구현하여 상대방 목소리가 친절하게 들리도록 하는 친절기능을 부가한 전화기를 구현한다.

1. 서 론

실생활에서 통신용으로 아주 널리 사용되고 있는 전화기의 기능을 개선하는 방법에 관한 것으로 상대방으로 걸려오는 전화의 목소리는 각양각색이다. 상대방이 보이지 않기 때문에 급한 목소리, 욕하는 소리, 사투리가 섞인 소리, 불명료한 목소리 등등으로 수신자의 감정을 불쾌하게 만든다. 이럴 때에 필요한 전화기가 바로 친절전화기인데 수신자가 전화기에 부착된 친절-보턴(또는 특정 키-보턴)을 누르면 상대방의 목소리가 친절하면서 자세한 목소리로 천천히 들리도록 구현한 것이다. 특히, 청각 장애인이나 노인층의 경우에는 청각기능이 저하되어 평균 발성속도로 이야기를 진행하여도 잘 알아듣지 못하는 경우에도 친절전화기의 기능은 발성을 천천히 또렷하게 들려줌으로서 복지통신 분야에 필수적인 기능으로 활용될 수 있다. 또한 불특정 다수의 고객을 전화 통신으로 영접하는 관련 서비스업 종사자들은 고객의 다양한 목소리의 형태로 인해 스트레스를 많이 받게 된다. 이러한 경우에도 친절전화기능은 고객의 목소리를 친절하고 차분하게 만들어 주기 때문에 목소리 관련 직업인들의 스트레스를 어느 정도 해소할 수 있다. 친절전화기는 전화기를 통해 수신되는 상대방의 목소리 정보를 분석하여 상대방의 개성정보는 그대로 두고, 의미를 나타내는 정보는 늘려 줌으로서 마치 동영상에서 슬로우-모션을 구현하는 것처럼, 목소리의 슬로우-오디오 기능을 구현한 것이다. 상대방의 목소리가 친절하게 들리도록 하기 위해서는 상대방의 실제 목소리 발성속도보다 천천히 들리게 해야 하는데, 처리된 데이터를 대기시키는데 필요한 메모리 버퍼를 링버퍼라고 한다. 따라서 본 논문에서는 지속시간을 조절하여 슬로우-목소리를 구현하거나, 발성 지속시간의 지연을 유성 및 비유성 구간으로 구분하여 처리를 다르게 하는 방법을 제안 한다.

2. PSOLA 알고리즘

PSOLA 알고리즘은 우선 피치주기 단위로 음성 파형을 분해한 다음 분해된 피치 단위에 윈도우 함수를 곱해서 단구간ST(Short-Term)신호의 열로 만들고 분해된 단위를 피치를 높일 때는 짧은 구간의 신호들을 재결합할 때 배치 간격을 넓히고, 피치를 낮출 때는 짧은 구간의 신호들의 배치 간격을 좁혀서 재결합한다. 그리고 합성음의 지속시간 조절은 짧은 구간 신호를 반복 삽입하여 지속 시간을 늘이거나 삭제하여 지속시간을 줄인다. 신호의 재결합은 합성음의 지속시간과 피치를 고려하여 짧은 구간의 신호를 재배치하여 중첩-가산한다.[1]

2-1. 피치 동기 분석과정

원래 음성 파형이 유성음인 경우에는 피치단위로 분해한 다음 윈도우 함수를 곱하여 ST신호의 열로 만든다. 무성음인 경우에는 10ms의 주기로 일정하게 분석한다. 분석 윈도우 함수에는 Hanning Window를 사용한다. 이런 윈도우 함수를 원래의 음성 샘플에 곱함으로써 다음 식(2.1)과 같은 피치 단위로 분해된 샘플열들을 얻는다.

$$S_{analysis}(n) = W_{analysis}(m-n)S(n) \qquad (2.1)$$

$S_{analysis}(n)$: 피치주기 단위의 ST 신호
$W_{analysis}(n)$: 분석 윈도우 함수
m : m번째 피치
$S(n)$: 원 음성 파형

2-2. 피치 스케일 변경

분석과정에서의 ST신호의 열은 원래의 음성 샘플의 피치단위로 배열되어있다. 따라서 피치를 변경하기 위해서는 이 간격들을 변경할 피치 간격들로 재배열하면 된다. 다음 식(2.2)는 피치가 변경된 신호를 나타낸 것이다.

$$S_{synthesis}(n) = S_{analysis}(n - m_a) \qquad (2.2)$$

$S_{synthesis}(n)$: 피치가 변경된 ST신호
m_a : 변경할 피치 간격

따라서 피치를 높일 때는 ST 신호의 간격을 작게 배열하고, 피치를 낮출 때는 ST신호의 간격을 크게 배열하면 된다. 하지만 이런 순차적인 배열사이에서 정확한 피치 동기화를 유지하는 것이 중요하다. 이렇게 재배열된 ST신호에서 겹쳐지는 부분을 더해주면 된다.

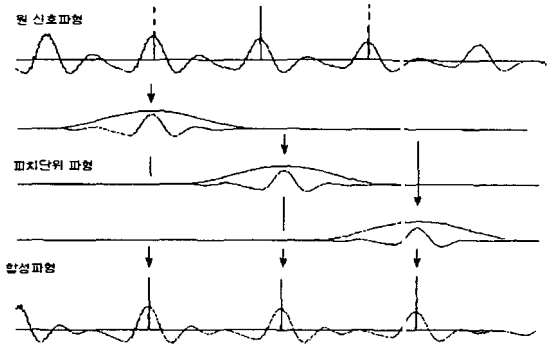

그림 2-1. PSOLA 합성방식에 의한 피치 변경의 예

본 논문에서는 선형예측분석을 이용한 피치시점 검출법을 이용하였고 그리고 위에서 구한 피치시점 정보를 이용하여 PSOLA 합성방법으로 피치를 변경하였다. 결과적으로 본 논문에서는 FFT변환 특성을 이용한 지속시간 변경법을 제안하고자 하며 또한 프레임처리에서 윈도우의 영향으로 스펙트럼 왜곡 및 음질 저하를 방지하기 위하여 보상된 윈도우를 적용하였다[1].

3. 제안한 지속시간 변경법

본 논문에서는 FFT변환 특성을 이용해 음색의 변경 없이 실시간으로 지속시간을 변경해 주는 방법에 대해 제안하고자 한다. 본 방법에서의 지속시간 변경법으로 FFT를 이용하여 계산시간을 줄이고 진폭과 위상에 각각 2^n배의 Interpolation과 Decimation을 수행한 다음 FFT point의 2^n배 point로 IFFT과정을 수행함으로써 스펙트럼의 변경 없이 지속시간을 변경하였다. 다음 그림 3-1은 시간-주파수 변환특성을 이용한 지속시간 변경법의 블록도를 나타낸 것이다.

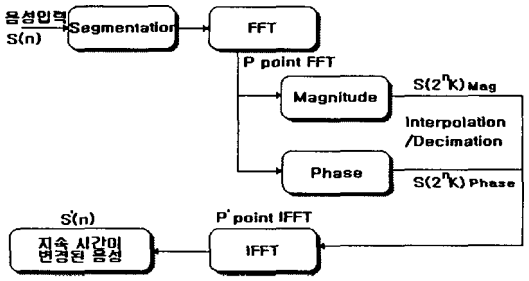

그림 3-1. 제안한 지속시간 변경법

그림 3-1의 블록도를 설명하면 우선 Frame 단위로 Segment된 음성신호를 FFT 통해 진폭성분과 위상성분으로 나눈다. 그런 다음 이렇게 얻어진 각 진폭과 위상성분에 2^n배로 주파수축 Interpolation과 Decimation과정을 수행한다. 그런 다음 2^n배 point로 IFFT를 수행하여 지속시간이 변경된 음성을 얻어낼 수 있었다. 또한 윈도우의 영향으로 인해 스펙트럼 왜곡 및 음질저하를 방지하고 보상하기 위하여 프레임 처리는 식(3.1)과 같다. 최소한의 에너지 누설을 막고 실험결과 해밍 윈도우보다 음질이 좋은 Rectangular 윈도우를 적용하였고 보상전보다 피크성분이 강조된 피치를 얻을 수 있었다[3].

$$S\_syn = S\_syn * (W(n)/R\_S\_syn) \qquad (3.1)$$

여기서 S_syn는 합성음 데이터, R_S_syn는 실 합성음 데이터이다. 그림 3-2는 한 프레임 내에서 본 논문에서 제안한 FFT변환 특성과 보상된 프레임처리를 이용하여 지속시간을 변경한 예에 대해 나타낸 것이다.

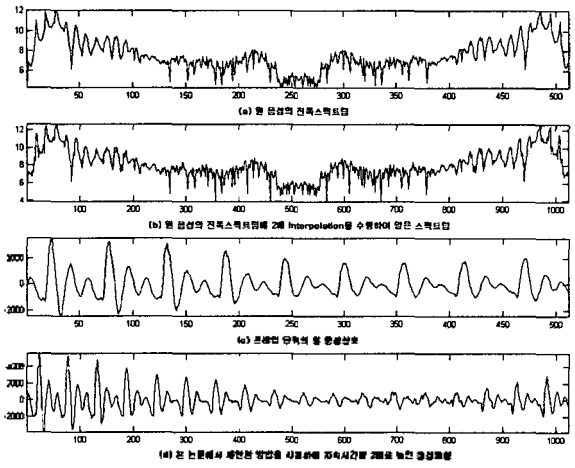

그림 3-2. 제안한 방법과 보상된 프레임 처리를 이용하여 지속시간을 변경한 예

4. 친절 전화기의 구현

4-1. 원리

소프트 사운드는 그림 4-1과 같이 기존 전화기에 내장된 컴퓨터 칩에서 목소리를 분석하여 발성자의 목소리특성을 그대로 유지하면서 발성시간이 길게 합성되도록 하는 첨단 처리기능을 추가한 것이다. 즉, 목소리는 성대의 떨림과 목구멍에서의 공명에 의해 소리가 발생하는데, 이러한 목소리의 생성원리를 이용하여 목소리의 특징은 그대로 두고 말하는 의미 정보만을 뽑아서 반복하여 합성하면 천천히 들리면서 명료하고 친절한 목소리로 바뀌게 된다. 친절전화기의 핵심기술은 사람의 목소리에서 말뜻을 나타내는 음운정보와 개성을 나타내는 운율정보를 자동으로 분류하여 개성을 보존하면서 동시에 음운정보를 지속함으로써 목소리의 친절성을 증대시켰다는 점이다.

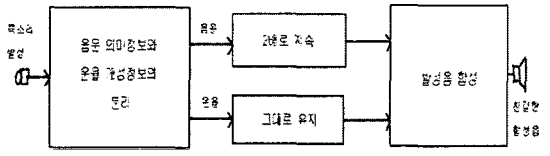

그림 4-1. 친절전화기의 블럭도

4-2. 하드웨어 처리

마이크로폰이나 전화라인 등으로부터 들어오는 아날로그 형태의 목소리 신호를 입력 받아서 친절한 목소리로 발성처리를 하는 장치는 그림 4-2와 같다. 아날로그 형태로 입력된 목소리 신호파형은 증폭기에서 증폭된 다음에 엘리어징(aliasing)효과를 제거하기 위해 저역통과여파기를 통과하고, 양자화(quantization) 및 부호화(coding)를 수행하는 아날로그-디지털 변환기를 통과함으로서 선형펄스부호변조(PCM) 형태의 디지털 신호로 바뀌어서 범용 CPU나 디지털 신호처리기(DSP)에서 소프트웨어나 펌웨어에 의해 처리된다. 신호처리 될 때는 이 컴퓨터 처리기가 대내외에 설치된 주변장치를 참고할 수도 있고, 또한 입력 디지털 신호나 처리 결과를 저장하기 위해 주변 메모리를 참고할 수도 있다. CPU에서 소프트웨어에 의해 발성변환 처리된 디지털 신호는 디지털-아날로그 변환기를 통해 표본화된 아날로그 신호형태로 변환된다. 이 신호를 저역통과 여파기에 통과시키면 양자화 잡음이 제거된 아날로그 신호가 되고, 적당히 증폭하면 전화 수화기나 스피커 등을 통해서 들을 수 있는 아날로그 신호가 된다.

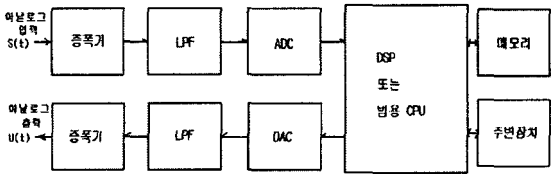

그림 4-2. 발성처리 구성도

4-3. 지속시간 버퍼링 제어

친절전화기는 기존 전화기의 기능을 수행하는 CPU칩에 친절기능의 소프트웨어나 펌웨어를 추가한 것이다. 그림 4-3은 소프트 사운드 전화기의 전체 흐름도이다.

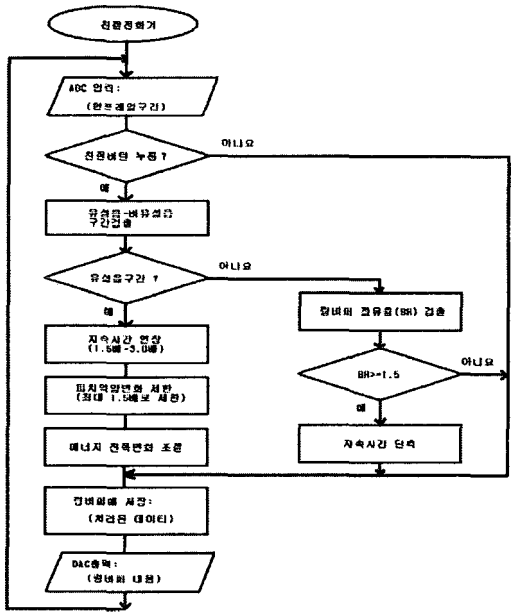

그림 4-3. 친절전화기의 흐름도

전화통화가 이루어 졌을 때에 친절버튼이 눌러지지 않았다면 기존 전화기와 같이 목소리 통신을 수행하게 된다. 친절전화기능이 시작되면 아날로그-디지털 변환기(ADC)에서 입력된 데이터 표본값이 한 프레임단위로 동시에 처리된다. 먼저 현재 프레임에 있는 데이터 값이 유성음 구간인지 아닌지를 파악하고, 유성음 구간이 아니면 링버퍼의 점유율(Buffer Rate, BR)을 계산하게

된다. 상대방의 목소리가 친절하게 들리도록 하기 위해서는 상대방의 실제 목소리 발성속도보다 천천히 들리게 해야 하는데, 처리된 데이터를 대기시키는데 필요한 메모리 버퍼를 링버퍼라고 한다.

링버퍼의 점유율(BR)은 친절기능에서 처리된 데이터가 링버퍼에서 대기되는 시간비율을 나타내는데, 현 프레임이 비유성음구간이고 링버퍼에 대기하고 있는 시간이 정해진 시간(예 BT=1.5이상)을 넘어섰다면, 발성속도를 앞당기도록 발성의 지속시간 단축을 수행하게 된다. 이렇게 함으로써 친절기능이 수행될 때 야기되는 발성시간 지연을 해소할 수 있게 된다. 즉, 유성음 구간에서는 친절하고 또렷하게 발성되도록 데이터를 천천히 출력하지만 비유성음 구간에서는 발성속도를 빠르게 하여 전체적인 시간지연을 해소하게 한 것이다.

현재의 프레임이 유성음 구간인지 비유성음 구간인지를 측정하는 방법은 음성처리 기술에서 많이 제안되어져 있으며, 일례로 에너지 레벨을 측정하여 쉽게 파악할 수 있다. 즉, 현재 프레임의 평균 에너지가 정해진 문턱 값 이하라면 이 구간은 비유성음 구간이 된다.
현재의 프레임의 데이터가 유성음 구간이라면 이 데이터에 대해 친절기능 처리를 수행하게 된다. 친절기능은 이 데이터의 발성속도를 천천히 지속하기 위해 지속시간(예, 1.5~3.0배 정도)을 연장시킨다. 유성음 데이터의 지속시간 변경은 피치주기 단위로 수행하였고, 이때 피치주기를 정확히 검출해야 한다. 또한 유성음 구간내에서 억양의 변화를 어느 정도로 제한(예, 1.5배 이내)하기 위해, 연속된 유성음 구간의 피치주기를 검출한 다음에 프레임당 변화도를 구하고, 변화가 크다면 피치주기변경을 수행하여 목소리를 안정시키게 된다.

피치주기의 변경은 피치주기 검출이 잘 이루어진 다음에 이를 근거로 피치주기를 변경시키게 된다. 이렇게 처리된 데이터들은 파형의 진폭이 자연스럽지 못하고 부자연스럽게 되므로 이를 진폭의 변화가 자연스럽게 이어지도록 하는 에너지 진폭변화 조절을 수행해야 한다. 일예로 에너지 진폭의 변경은 피치주기 단위로 처리하며, 한 피치주기의 평균 에너지 진폭을 곱함으로서 수행한다. 이렇게 처리 완료된 음성 데이터들은 링버퍼에 저장시키고 저장된 순서에 따라서 디지털-아날로그 변환기(DAC)를 통해 음성 데이터 표본 단위로 수화기나 스피커폰을 통해 출력한다. 여기서 친절전화기의 기능은 실시간으로 처리된다.

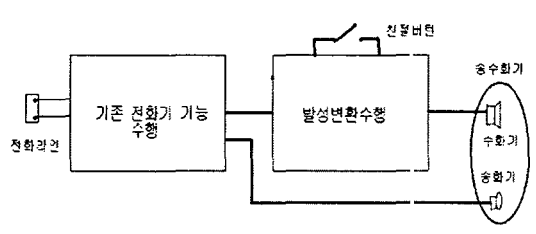

그림 4-4. 친절전화기의 구성도

즉, 아날로그-디지털 변환기(ADC)에서 한 프레임의 데이터를 받고나서부터 그 다음 프레임의 데이터를 받아올 때까지 친절전화기능의 처리가 끝날 수 있도록 해야만 한다. 그림 4-4는 본 논문에서 제안한 구성도이다.

5. 결 론

본 논문에서 제안한 방법은 전화기의 수화기에서 들리는 상대방의 목소리를 디지털 발성처리기술을 적용하여 천천히 친절하게 들리도록 하는 통화하는 방식을 새로이 제안하는 것이다. 목소리 신호의 특징 추출을 수행하여 발성자의 특성정보는 그대로 유지하면서 발성자의 의미정보를 친절하게 변경하는 것으로서, 발성자의 발성특성에서 지속시간을 조절하여 슬로우-목소리를 구현하거나, 발성 지속시간의 지연을 유성 및 비유성 구간으로 구분하여 처리를 다르게 하는 등의 발성변환법을 전화기에 구현하여 상대방 목소리가 친절하게 들리도록 하는 친절기능을 부가한 전화기를 구현하였다.
더불어 본 논문에서는 FFT 변환특성을 이용한 주파수 영역에서의 지속시간 변경법으로 지속시간을 변경하였다. 또한 프레임처리에서 윈도우의 영향으로 스펙트럼 왜곡 및 음질 저하를 방지하기 위하여 보상된 윈도우를 적용하였다. 만약 운율조절에 있어서 지속시간을 자유롭게 변경할 수 있다면 언어장애인의 발음교정이나 어학학습 등 여러 분야에 이용할 수 있을 것이다[5].

감사의 글

본 연구는 한국과학재단 특정기초연구(과제번호 R01-2002-000-00278-0)의 지원에 의하여 이루어 졌음.

6. 참 고 문 헌

[1] G. Bristow, *Electronic Speech Synthesis*, McGraw-Hill, 1984.

[2] J.R. Deller, J.G. Proakis, J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Co., 1993.

[3] 박형빈, 조왕래, 김종득, 박원, 심도식, 배명진, "피치변경율에 따른 최적의 피치 변경법에 관한 연구", 제15회 음성 통신 및 신호처리 워크샵 논문집, Vol.15, No.1, PP.460-464, 1998년 08월 21-22일.

[4] B.E. Caspers and B.S. Atal, "Changing Pitch and Duration in LPC Synthesised Speech using Multipulse Excitation." J. Acoust. Soc. Amer., Vol.73, No.1, pp.55, 1983.

[5] 하정호, 정재호, "합성음 구현을 위한 음의 억양과 장단의 변환 연구", 제 11회 음성통신 및 신호처리 워크샵 논문집, 제 SCAS-11권 1호, pp.328-333, 1994년 10월 28일.