한국퍼지 및 지능 시스템학회 논문지 1998, Vol. 8, No. 2. (1998)60~69

고음질의 음성합성을 위한 퍼지벡터양자화의 퍼지니스 파라메타선정에 관한 연구

A Study on Fuzziness Parameter Selection in Fuzzy Vector Quantization for High Quality Speech Synthesis

이 진 이
Jin-Yi Lee
충남산업대학교 전자공학과

요 약

본 논문에서는 퍼지 벡터양자화를 이용하여 음성을 합성하는 방법을 제시하고, 원음에 가까운 합성음을 얻기 위하여 퍼지벡터양자화의 성능을 최적화 하는 Fuzziness값의 선정방법을 연구한다. 퍼지벡터 양자화를 이용하여 음성을 합성할때, 분석단에서는 입력 음성패턴과 코드북의 음성패턴의 유사도를 나타내는 퍼지 소속함수값을 출력하고, 합성단에서는 분석단에서 얻은 퍼지소속 함수값, fuzziness값, 그리고 FCM(Fuzzy-C-Means) 연산식을 이용하여 음성을 합성한다. 시뮬레이션을 통하여 벡터양자화에 의해 합성된 음성과 퍼지 벡터양자화에 의해 합성된 음성을 코드북의 크기에 따라 비교한 결과, 퍼지 벡터양자화를 이용한 음성합성의 성능이 코드북 크기가 절반으로 줄어도 벡터양자화에 의한 성능과 거의 같음을 알 수 있다. 이것은 VQ(Vector Quantiz-ation)에 의한 음성합성 결과와 같은 성능을 얻기 위해서 퍼지 VQ를 사용하면, 코드북 저장을 위한 메모리의 크기를 절반으로 줄일 수 있음을 의미한다. 그리고 SQNR을 최대로 하는 퍼지 벡터양자화를 얻기 위한 최적 Fuzziness값은 음성분석 프레임의 분산값이 크면 작게 선정해야 하고, 작으면 크게 선정 해야함을 밝혔다. 또한 합성음들을 주파수 영역의 스펙트로그램에서 비교한 결과, 포만트 주파수와 피치 주파수에서 퍼지 VQ에 의한 합성음성이 VQ에 의한 것보다 원 음성에 더 가까움을 알 수 있었다.

ABSTRACT

This paper proposes a speech synthesis method using Fuzzy VQ, and then study how to make choice of fuzziness value which optimizes (controls) the performance of FVQ in order to obtain the synthesized speech which is closer to the original speech. When FVQ is used to synthesize a speech, analysis stage generates membership function values which represents the degree to which an input speech pattern matches each speech patterns in codebook, and synthesis stage reproduces a synthesized speech, using membership function values which is obtained in analysis stage, fuzziness value, and fuzzy-c-means operation. By comparsion of the performance of the FVQ and VQ synthesizer with simmulation, we show that, although the FVQ codebook size is half of a VQ codebook size, the performance of FVQ is almost equal to that of VQ. This results imply that, when Fuzzy VQ is used to obtain the same performance with that of VQ in speech synthesis, we can reduce by half of memory size at a codebook storage. And then we have found that, for the optimized FVQ with maximum SQNR in synthesized speech, the fuzziness value should be small when the variance of analysis frame is relatively large, while fuzziness value should be large, when it is small. As a results of comparsion of the speeches synthesized by VQ and FVQ in their spectrogram of frequency domain, we have found that spectrum bands(formant frequency and pitch frequency) of FVQ synthesized speech are closer to the original speech than those using VQ.

1. 서론

음성신호는 사람만이 갖는 의사전달의 최고수단이며, 상황에 따라 변하는 정보의 정확한 전달을

위해서 고도의 지적능력을 필요로 하기 때문에 음성언어는 단순하지 않는 특수한 구조를 갖고 있다. 이처럼 결코 단순하지 않은 인간의 음성언어를 인간에 국한하지 않고 인간과 기계사이의 의사전달을 가능하게 하기 위한 소위 인간과 기계간의 인터페이스에 관한 음성공학의 연구가 활발히 진행되고 있다. 국내에서는 문장을 음성으로 변환하여 합성하는 TTS(text-to-speech)의 상품화를 시도하였다. 그러나 수 메가에서 수십 메가의 음성 DB는 음성합성기의 휴대화에 장애가 되고 있고, 음성데이타의 수집 및 압축(코드북화)은 음성합성 연구자가 해결해야 할 문제이다. 음성처리는 크게 다음으로 분류된다. 주로 인간과 인간사이의 정보전달을 다루는 음성부호화와 전송분야, 기계와 인간사이의 정보전달을 다루는 음성합성분야, 인간과 기계간의 정보전달을 다루는 음성인식분야로 분류할 수 있으며, 이들을 위해 기본적으로 수행되는 연구분야로는 음성발생, 음성분석, 음성이해, 음성평가 등으로 광범위하다[1].

본 연구는 이러한 음성처리의 여러 분야중 음성합성에 관한 내용으로 음성합성 방법에는 크게 3가지로 구분할 수 있다. 첫째, 파형 부호화 방법인데, 이 방법은 파형을 부호화하여 저장한 다음 원하는 음성을 재생하는 방법이다. 이 방식은 다른 방식에 비하여 매우 간단하며 대표적인 방식은 PCM으로 64kbps에서 좋은 음질을 갖는다. 한편 PCM 방식과 같이 디지탈 음성파형을 그대로 저장하기 보다는 음성 샘플간의 상관성을 이용한 예측 잔차를 부호화하여 저장하는 예측 부호화방식[2][3]이 있다. ADPCM과 ADM이 대표적이다. 둘째, 분석-합성방법인데, 이 방식은 음성을 분석하여 음성이 내포하고 있는 특징만을 추출해서 부호화하는 방식으로 LPC[4]나 Formant 방식이 대표적이다. 이 방식은 파형부호화 방법보다 훨씬 정보량이 줄어 들어 8Kbps 이하로 매우 낮으나, 그 구조가 매우 복잡하고 음질도 다른 방법에 비해 많이 떨어지는 편이다. 세째는 혼합 부호화방식이다. 이 방식은 펄스와 잡음을 음원으로 하는 분석-합성방법과 파형 부호화방식을 결합한 방식으로 MPLPC(Multi Pluse LPC), RELP(Residual Excited LPC), VELP(Voice Excited LPC), CELP(Code Excited LPC)와 같은 시간축상의 방식과 ATC(Adaptive Transform Coding), SBC(Subband Coding)과 같은 주파수 영역상의 방식이 있다. 이러한 방식은 저주파 성분의 음성이 정확하게 재생되며, LPC 분석방법에 의한 것 보다 모든 주파수 대역들의 스펙트럼이 정확히 표현되고, 아울러 피치추출과 유성/무성음의 판단이 불필요하며 잡음에 대해서도 강하다. 그러나 시스템이 복잡하고 계산량이 많은 것이 단점이지만 파형부호화방식보다 데이타량도 훨씬 적고 자연성과 명료도가 분석-합성방법보다 훨씬 우수하다.

음성신호를 스칼라방식을 이용하여 디지탈화할 경우 비트율이 커지고 대역폭도 증가되는 것이 일반적인 문제점이다. 이러한 점에서 음성신호의 허용 가능한 충실도와 질을 유지하면서 방대한 양의 음성데이타를 저장 또는 전송하기 위해서 데이타 압축이 필요하다. 이를 위해 샘플링한 음성데이타를 하나의 블럭으로 묶어 벡터 처리하는 벡터양자화(Vector Quantization ; VQ)[5]가 연구되었다. VQ는 음성부호화, 음성인식, 패턴인식 등에 많이 응용되고 있다. 이러한 VQ의 성능은 코드북에 크게 의존하며, 코드북 작성에는 K-means 알고리즘, LBG 알고리즘, 퍼지군집화 알고리즘[6], 경쟁학습 신경망 알고리즘[7] 등이 있다. 일반적으로 VQ의 성능은 코드북의 크기에 의존하는데, 본 논문에서는 코드북 크기에 따른 메모리를 늘이지 않고도 기존의 VQ에 의한 성능보다 우수한 기능을 갖는 퍼지벡터양자화를 이용하는 음성합성방식을 제시한다. 이 방식은 합성 단위간의 연속부에서 발생하는 불연속성을 제거하기 위해 퍼지이론을 적용한 것이다.

퍼지이론은 퍼지집합, 퍼지논리, 퍼지측도의 3가지 틀로 구성된다. 퍼지집합은 어떤집합에 속하는가 속하지 않는가를 0과 1로 나타내는 기존의 논리체계(hard decision)와는 달리 어느 정도 속하는가 아니면 어느 정도 속하지 않는가 하는 애매한 논리체계(soft decision)를 0 과 1 사이의 소속정도로 표현하여 처리한다. 이러한 개념을 VQ에 적용한 것이 퍼지벡터양자화[8]이다. 퍼지벡터양자화를 음성인식과 단어인식[9]에 이용한 연구가 있다.퍼지벡터양자화의 분석단에서 입력 음성벡터와 코드벡터 사이의 퍼지판정을 수행하여, 입력벡터에 대한 각 코드벡터와의 소속정도를 나타내는 소속함수값을 얻는다. 그리고 합성단에서는 Fuzzy c-means연산식을 사용하여 합성한다.

음성합성에 관한 연구의 방향은 적은 음성데이타로 고 품질의 합성음질을 얻는 것에 있기 때문에, 본 연구에서는 퍼지이론을 적용하여 메모리의 양을 줄이면서 합성음질을 높일 수 있는 방법을 위해 퍼지벡터양자화를 음성합성분야에 적용하여 코드북의 크기에 따른 메모리를 줄이고 기존의 VQ 방식에 의한 것보다 합성음질을 높히는 방식을 제안한다. 2장에서는 코드북 작성에 대한 몇가지 알고리즘을 정리하고, 3장에서는 퍼지벡터양자화를 이용한 음성분석-합성시스템에 대해 기술한다. 4 장에서는 K-means 알고리즘으로 작성된 주어진 코드북 크기에 대해 VQ와 퍼지벡터양자화에 의한 음성합성의 결과를 비교하고, 원 음성을 최적으로 합성할 수 있는 최적 퍼지 벡터양자화를 위한 Fuzziness값 설정에 대해 기술하고 5장에서 결론을 맺는다.

2. 코드북 설계 알고리즘

코드북을 작성하는 방법은 여러가지가 있다. K-means 알고리즘 과 LBG 알고리즘[10]은 먼저 입력벡터와 같은 차원을 갖는 L-레벨 초기 코드북을 작성한 다음, 계산된 평균 왜곡변화의 임계값이 특정값 이하가 될 때 까지 반복하여 갱신함으로써 만들어진다. 이 두 방법의 차이점은 K-means 알고리즘은 코드북의 크기가 처음에 정해지는 반면에, LBG 알고리즘은 전체 평균왜곡이 규정치 이하가 될 때 까지 코드북 크기를 늘여 나가면서 원하는 크기의 코드북을 작성하는 방식으로 K-means 알고리즘 보다 규정 왜곡치로 향한 수렴성이 빠르다. 이러한 방법은 명확한 경계선을 넣어서 데이터의 집합을 몇 개의 집단으로 분류하는 것이지만, 입력벡터가 군집의 중심벡터에 소속하는 정도를 계산하여 입력벡터들을 군집화하는 기술이 있는데, 이것이 FCM 군집화(clustering) 방법[11] 이다. 그러나 이 방법은 많은 계산량을 필요로 하는 단점을 갖고 있다. 그 외 신경망의 간단한 학습기능을 이용하여 코드북을 작성하는 방법[12, 13]이 있는데, 대표적으로 경쟁 학습신경망 방법에 의한 승자뉴런(winner neuron) 하나 만을 학습시키는 알고리즘이 있다.

2.1 K-means 알고리즘

이 알고리즘은 먼저 입력벡터와 같은 차원의 L 레벨 초기 코드북을 작성한 다음 임계 평균왜곡치가 주어진 값 이하가 될 때 까지 코드북을 반복하여 갱신함으로써 만들어진다. 이렇게하여 얻어진 최종 코드북은 센트로이드(centroid)라고 하는 중심벡터인 코드벡터로 구성되며 작성 절차는 다음과 같다. 입력벡터와 코드벡터와의 거리는 음성부호화의 왜곡 측정 방법인 평균자승오차방법을 사용한다.

단계 1) 초기화단계

L (코드북 레벨), ε (평균왜곡 D 변화의 임계값), m (반복횟수), D(-1) (초기 전체 평균왜곡), n (입력벡터 수)를 정한다.

단계 2) 초기 코드북 작성단계

초기 코드벡터 $(Y_j(0), 1 \le j \le L)$의 값을 설정한다. 입력벡터성분들의 최대크기와 최소크 기를 고려하여 랜덤하게 설정하거나, 입력벡터의 첫번째 값으로 설정하여 L개의 초기 코드 벡터를 갖는 코드북을 작성한다.

단계 3) 입력벡터 $\{X_i, 1 \le i \le n\}$가 다음 식을 만족하면 집단 C_j로 분류

$$X_i = C_j \quad if \quad d\left[X_i, Y_j(m)\right] \le d\left[X_i, Y_k(m)\right] \quad all \quad j \ne k, \quad 1 \le j \le L \tag{1}$$

단계 4) 각 집단의 중심벡터를 계산하여 코드벡터를 다시 선정

$$Y_j(m) = Cent(C_j(m)), \quad 1 \le j \le L \tag{2}$$

단계 5) 전체 평균왜곡 D(m)을 계산

$$D(m) = E\left[d(X_i, Y_j(m))\right] \tag{3}$$

단계 6) 다음 (4)식을 만족하면 수행을 멈추고 만족하지 않으면 단계 3)을 수행한다.

$$\frac{D(m-1) - D(m)}{D(m)} \le \varepsilon \tag{4}$$

2.2 FCM 군집화(clustering) 알고리즘

퍼지군집화(fuzzy clustering) 방법은 통계적 기법으로써 단순군집화(crisp clustering)방법(K-means와 LBG 알고리즘)을 확장한 개념이다. 단순 군집화방법은 명확한 경계선을 넣어서 데이터집합을 몇 개의 군집으로 분류하는 것이지만 실제로 군집들의 경계부에 존재하는 데이터를 어느 한 개의 군집에만 완전히 소속시키기에는 무리가 있다. 그래서 소속함수값을 {0,1}의 두 개의 값에서 [0,1]로 확장해서 경계부의 데이터를 여러 군집에 약간씩 소속하게 한 것이 퍼지군집화기법이다. 이 알고리즘에 의한 코드북 작성의 과정은 다음과 같다.

단계 1) 초기화단계

결정된 L개의 군집의 초기 중심벡터 Y_j $(1 \le j \le L)$를 설정하고 모든 입력벡터가 L개의 군집의 중심벡터에 소속되는 정도인 소속함수값을 1/L로 초기화하다. 소속함수값의 변화량 δ를 설정한다.

단계 2) 소속함수값 계산

모든 입력벡터들이 모든 군집의 중심벡터에 소속되는 소속정도를 다음 식으로 새로 계산한다.

$$\mu_{ij}(new) = \left\{ \sum_{k=1}^{L} \left[\frac{d(X_i, Y_j)}{d(X_i, Y_k)} \right]^{\frac{1}{F-1}} \right\}^{-1} \qquad (5)$$

여기서 μ_{ij}는 i 번째 입력벡터가 j 번째 군집의 중심벡터에 소속되는 값 (소속함수값) 이고 $d(X_i, Y_j)$는 벡터의 거리 (음성부호화의 왜곡 측정값인 평균자승오차)를 나타낸다.

단계 3) $\max\left[\left|\mu_{ij} - \mu_{ij}(new)\right|\right] < \delta$ 이면 멈춘다. 그렇지 않으면 단계 4) 로 간다. 여기서 $0 < \delta < 1$이다.

단계 4) 새로운 중심벡터 Y_j(new)를 다음 식으로 구한다.

$$Y_j(new) = \frac{\sum_{i=1}^{n} \left[(\mu_{ij})^F X_i \right]}{\sum_{i=1}^{n} \left[(\mu_{ij})^F \right]} \qquad (6)$$

여기서 i=1,...,n (입력벡터의 수), j=1,...,L (중심벡터의 수 = 코드북 레벨)을 나타낸다.

단계 5) 단계 2 를 수행한다.

2.3 경쟁학습신경망 알고리즘

경쟁 학습신경망의 구조는 입력층, 은닉층, 출력층으로 구성되며 입력층에서는 입력벡터를 받아들인다. 어떤 벡터 X_i가 주어졌다면 X_i의 모든 성분들은 입력층 뉴런에 일괄적으로 입력되며, 은닉층에서는 입력벡터와 연결강도(입력층과 은닉층뉴런 사이의 연결강도) 사이의 거리를 계산한다. 출력층에서는 은닉층에서 계산한 거리를 비교하여 최소거리의 뉴런을 택하여 승자뉴런으로 결정한다. 이 승자뉴런만 학습하여 학습이 종료되면 입력층과 은닉층의 연결강도가 코드벡터를 형성하여 L 레벨의 코드북이 작성된다. 즉

$Y=[Y_j]_{1 \le j \le L}$

단계 1) 초기화 단계

입력벡터의 차원(코드벡터의 차원)과 은닉층 뉴런 수(코드북 크기)L, 학습률 $\varepsilon(n)$ 결정

단계 2) 입력층과 은닉층사이의 연결강도($Y_j(0)$)의 초기화

랜덤하게 선택하거나 입력데이타 집합의 첫 번째 벡터로 초기화

단계 3) 받아들인 입력벡터와 연결강도 (입력층과 은닉층사이의 연결강도)의 거리를 계산하여 가장 적은 거리값을 갖는 승자뉴런을 결정한다.

$$Z_j = \begin{cases} 1, & if \quad d(X_i, Y_j(m)) \le d(X_i, Y_k(m), j \ne k, 1 \le j \le L \\ 0, & otherwise \end{cases} \qquad (7)$$

단계 4) 결정된 승자뉴런을 다음 식으로 학습한다.

$$Y_j(m+1) = Y_j(m) + \varepsilon(m)(X_i - Y_j(m))Z_j \qquad (8)$$

$$\varepsilon(m) = Ae^{(-m/T)} \qquad (9)$$

단계 5) 모든 입력벡터에 대한 군집당 중심벡터와의 평균거리가 수렴하거나 원하는 값에 도달 하면 훈련을 끝내고 그렇지 않으면 단계 3) 부터 계속 수행한다.

여기서, d는 벡터의 거리, $\varepsilon(m)$은 학습률을 나타내며 입력벡터의 패턴들이 군집(cluster)의 중심벡터로 접근함에 따라 그 변화량이 시간에 따라 감소하는 지수함수형을 취한다. A는 $\varepsilon(m)$의 최대 변

화량을 결정하는 상수이고, T (상수)와 함께 입력 데이타의 통계적 특성을 고려하여 최적의 값으로 선택된다.

3. 퍼지벡터양자화 음성-분석 합성 시스템

벡터양자화(VQ : Vector Quantization)는 입력벡터와 가장 유사한 코드벡터를 하나 발생시키는 반면에, 퍼지벡터양자화(Fuzzy VQ)는 하나의 입력벡터에 대해 코드북의 모든 코드벡터와의 유사정도를 0 과 1 사이의 값으로 수치화한 소속함수값을 성분으로 하는 하나의 벡터(확률질량벡터)를 발생한다. 합성단에서는 이들 소속함수값과 합성규칙(FCM 연산식)을 사용하여 음성을 합성한다. FVQ의 이러한 양상은 VQ보다 입력벡터에 더욱 근접한 합성벡터를 발생하여 합성음질을 높일 수 있다.

3.1 퍼지벡터양자화의 음성분석

퍼지벡터양자화의 코드북은 기존의 VQ 코드북과 동일하며 퍼지 VQ 분석단에서는 퍼지집합이론을 적용하여 입력음성벡터와 각각의 음성코드벡터와의 소속함수값을 다음 식에 의하여 계산한다.

$$\mu_{ij} = \left\{ \sum_{k=1}^{L} \left[\frac{d(X_i, Y_j)}{d(X_i, Y_k)} \right]^{\frac{1}{F-1}} \right\}^{-1} \qquad (10)$$

여기서 첨자i 는 입력음성벡터의 순번을 나타낸다. 첨자j 와 k 는 코드북 내의 음성 코드벡터의 순번을 나타내며 각각의 최대값은 코드북 크기인 L 이다. 확률질량벡터 O_i 는 위 식에서 얻은 소속함수값을 성분으로 한다.

$$O_i = [\mu_{i1}, \mu_{i2}, \ldots\ldots\ldots\ldots, \mu_{iL}] \qquad (11)$$

분석벡터 O_i 성분들은 각각 양의 값이고 모두 합하면 1 이 된다. 그리고 F(Fuzziness)>1 는 확률적 변화량을 나타내는 애매성 상수이다. 확률질량벡터 O_i의 성분들 μ_{ij}는 다음 식의 퍼지목적함수를 최소화하도록 정의된다[14].

$$\min Z(\mu, Y) = \sum_{i=1}^{n} \sum_{j=1}^{L} (\mu_{ij})^F d(X_i, Y_j) \qquad (12)$$

분석벡터O_i 성분중 가장 큰 소속함수값의 성분에 해당하는 음성 코드벡터가 입력 음성벡터와 오차가 가장 작은 벡터를 나타내며 VQ의 합성코드벡터에 해당한다. 그래서 분석벡터 O_i 의 성분들은 입력 음성벡터 X_i 가 음성 코드벡터 Y_j로 각각 합성될 확률을 나타낸다.

$$X_i = [x_{i1}, x_{i2}, \ldots, x_{iL}] \rightarrow \begin{bmatrix} Y_1 = [y_{11}, y_{12}, \ldots, y_{1N}] \\ Y_2 = [y_{21}, y_{22}, \ldots, y_{2N}] \\ \ldots\ldots\ldots\ldots \\ Y_L = [y_{L1}, y_{L2}, \ldots, y_{LN}] \end{bmatrix} \rightarrow O_i$$

그림 3-1. 퍼지 VQ의 분석단
Fig. 3-1. FVQ speech analysis

Fuzziness값인 F가 무한히 커짐에 따라서 O_i의 모든 성분들은 1/L에 근접한 값들을 갖게 되고, F가 1 에 가까운 값을 갖을 수록 어느 하나의 성분만 1에 가까운 값을 갖게 되어 다른 모든 성분들은 0의 값으로 근접한다. 그래서 퍼지 VQ의 판정을 단순판정 (F->1) 으로 할 것인가, 아니면 매우 퍼지판정 (F->∞) 으로 할 것인가 하는 것은 F 값의 적절한 선택에 달려 있다.

3.2 퍼지벡터양자화의 음성합성

퍼지 VQ 음성합성은 VQ의 패턴매칭에 의한 것과 다르게 분석단에서 얻은 분석벡터 O_i 와 합성규칙을 사용하여 코드북내의 음성 코드벡터가 아닌 새로운 하나의 합성 음성벡터 X_i를 얻는다. 합성규칙은 FCM 연산식을 사용하는데, 이 연산식은 압축된 모든 음성 코드벡터를 고려하여 소속함수값에 따른 기여도를 반영하여 코드북에 있는 코드벡터가 아닌 새로운 합성 음성벡터를 얻음으로써 VQ의 합성음 보다는 원음에 더 가까운 합성음을 얻는다.

$$O_i = [\mu_{i1}, \mu_{i2}, \ldots\ldots\ldots\ldots, \mu_{iL}] \rightarrow \begin{bmatrix} Y_1 = [y_{11}, y_{12}, \ldots, y_{1N}] \\ Y_2 = [y_{21}, y_{22}, \ldots, y_{2N}] \\ \ldots\ldots\ldots\ldots \\ Y_L = [y_{L1}, y_{L2}, \ldots, y_{LN}] \end{bmatrix} \rightarrow O_i$$

그림 3-2. 퍼지 VQ의 합성단
Fig. 3-2. FVQ speech synthesis

여기서 합성음성벡터 $\hat{X}_i$는 다음과 같이 주어진다.

$$\hat{X}_i = [\hat{x}_{i1}, \hat{x}_{i2}, \hat{x}_{i3}, \ldots\ldots, \hat{x}_{iL}] \qquad (13)$$

$\hat{X}_i$ 의 성분, $\hat{X}_{ij}$ 는 다음의 FCM 연산식[14] 으로 구한다.

$$\hat{X}_{ij} = \frac{\sum_{j=1}^{L}[(\mu_{ij})^F x_{ij}]}{\sum_{j=1}^{L}[(\mu_{ij})^F]} \qquad (14)$$

($1 \leq i \leq n$, n은 입력벡터의 갯수)

$\hat{X}_{ij}$ 은 퍼지집합이론에서 추론벡터성분에 해당하며, 합성 음성벡터 $\hat{X}_i$ 의 성분들, 즉 합성 샘플값들을 나타낸다.

4. 실험결과[15]및 검토

퍼지 VQ 음성 합성방식의 성능을 평가하기 위해 사용된 음성 데이타는 문장 /안녕하십니까/ 이다. 이 파형을 그림 4-1 에 나타내었다. 이 음성데이타는 Ariel 사의 speech station 3.0 음성

패키지[16, 17, 18]를 사용하여 얻은 것이다. 이 시스템은 PC-56D DSP Coprocessor Board를 사용하며 DSP 56001, zero state 16 KRAM, 14 bit analog I/O signal channel, 24 bit parallel interface등 으로 구성되어 있다. 샘플링 주파수는 8 KHz 로 하였고 이득은 2배로 하였다. 발성주기는 1초 이며, 8000개의 샘플값이다.

성능지수는 신호대 양자화잡음비(SQNR), 합성음성파형, 스펙트로그램으로 평가한다. SQNR 은 다음으로 정의한다.

$$SQNR = 10LOG_{10}\frac{E[S^2]}{[E(S-\hat{S})^2]} \quad [dB]$$

여기서 S는 원 음성의 샘플크기이고 $\hat{S}$ 는 합성된 음성의 샘플크기이다.

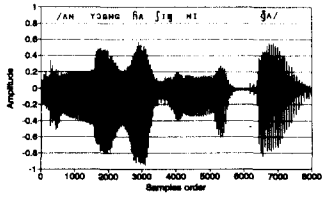


그림 4-1. 원 음성/안녕하십니까/

Fig. 4-1. Original speech /an young ha sim ni ka /

그림 4-2 (a), (b), (c), (d) 는 퍼지벡터양자화에 의한 최적의 합성음을 얻기 위해 Fuzziness 값에 따른 프레임 별 SQNR을 나타낸다. 음성의 분석 프레임은 400 샘플로 구성하였다. 사용된 코드북은 K-means알고리즘으로 작성한 4차원 32 레벨이다.

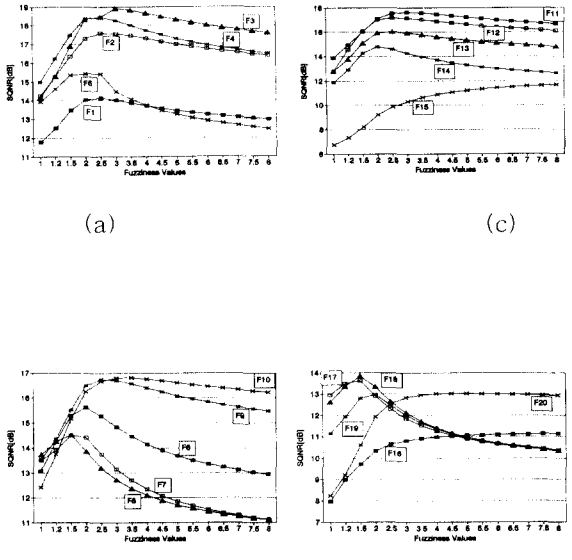


그림 4-2. 프레임 별 Fuzziness 값에 따른 SQNR

(a) 프레임 1-5, (b) 프레임 6-10, (c) 프레임 11-15, (d) 프레임 16-20,

Fig. 4-2. SQNR vs. Fuzziness values for frames order

(a) frame 1-5, (b) frame 6-10, (c) frame 11-15, (d) frame 16-20,

그림 4-2 (a)에서 Fuzziness값은 1.5에서 3의 범위에서, (b)에서는 1.5에서 3의 범위에서, SQNR을 최대로 함을 볼 수 있다. (c)에서는 Fuzziness값이 2에서 3의 범위에서 SQNR을 최대로 하지만 15번째 프레임에서는 8이상의 값에서 SQNR을 최대로 함을 알 수 있다. (d)에서는 Fuzziness값이 1.5에서 4의 범위에서 SQNR을 최대를 하지만 16번째 프레임에서는 8이상의 값에서 최대임을 볼 수 있

다. 음성의 분석프레임에 따라 퍼지벡터양자화에 의한 합성음의 SQNR을 최대로 하는 Fuzziness값이 서로 다름을 알 수 있다. 그래서 본 연구에서는 SQNR을 최대로 하는 최적 Fuzziness값을 정하기 위한 방법을 찾는 데 있다.

그림 4-3 에서는 프레임 별 분산값과 SQNR을 최대로 하는 Fuzziness값과의 관계를 나타낸 것으로 분산값이 큰 프레임에서는 작은 Fuzziness값을 택함으로써 SQNR을 높일 수 있음을 보여준다. 이러한 사실은 분산값이 큰 프레임에서는 Fuzziness값을 작게 부여함으로써 음성 코드벡터간의 상관성을 줄여 합성벡터를 얻음으로써 합성음의 음질이 향상되고, 분산값이 작은 프레임에서는 Fuzziness값을 크게 부여함으로써 음성 코드벡터간의 상관성을 크게 고려함으로써 음질을 향상시키게 된다.

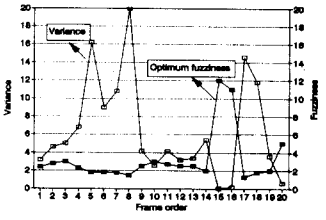


그림 4-3. 프레임별 분산값과 최적 Fuzziness값
Fig. 4-3. Variance and optimum fuzziness values for frames order

그림 4-4 는 32 레벨 VQ, 64 레벨 VQ, 32레벨 FVQ (Fuzzy VQ)에 의한 합성음성의 프레임당 SQNR을 나타낸다. 최적 Fuzziness값을 갖는 32 레벨 FVQ 의 성능은 64레벨 VQ 의 성능과 비교할 만함을 보여준다. 이는 코드북의 메모리를 절반으로 줄이고도 같은 합성음질을 얻을 수 있음을 보여준다.

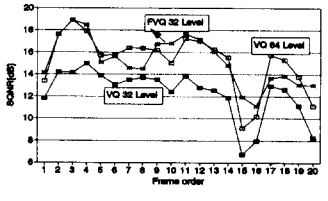
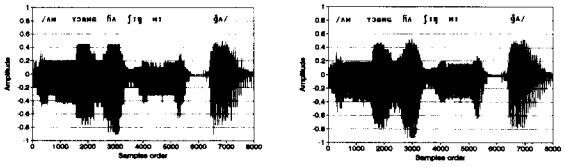


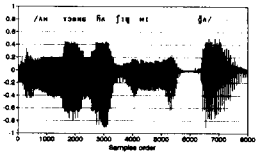
그림 4-4. VQ와 FVQ에 의한 합성음성의 프레임별 SQNR (a) 32 레벨 VQ, (b) 최적 퍼지니스를 갖는 32 레벨 FVQ, (c) 64 레벨 VQ

Fig. 4-4. SQNR vs. frames order for VQ using 32 level codebook (a), FVQ using 32 level codebook with optimum fuzziness values (b), and VQ 64 level codebook (c).

그림 4-5 (a), (b), (c) 는 각각 32 레벨 VQ, 64 레벨 VQ, 그리고 프레임별 최적 퍼지니스값을 갖는 32 레벨 FVQ에 의한 합성음성의 파형을 나타낸다. 코드북의 크기는 32 레벨로 같은 크기이지만 VQ방식 보다는 FVQ 방식에 의한 합성음의 포락선이 원음에 더 가까움을 알 수 있다.



(a) (b)

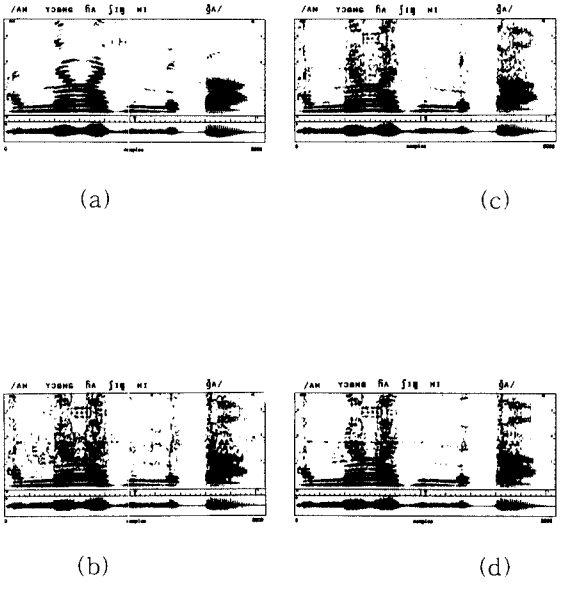


(c)

그림 4-5. 합성된 음성파형
(a) 32 레벨 VQ, (b) 64 레벨 VQ, (c) 32 레벨 FVQ

Fig. 4-5. (a) Speech synthesized by 32 level-VQ (b) 64 level-VQ (c) and 32 level-FVQ

그림 4-6 (a), (b), (c), (d) 는 각각 원 음성, 32 레벨 VQ, 64 레벨 VQ, 그리고 32 레벨 FVQ 에 의한 합성음성의 스펙트로그램을 나타낸다.



(a) (c)

(b) (d)

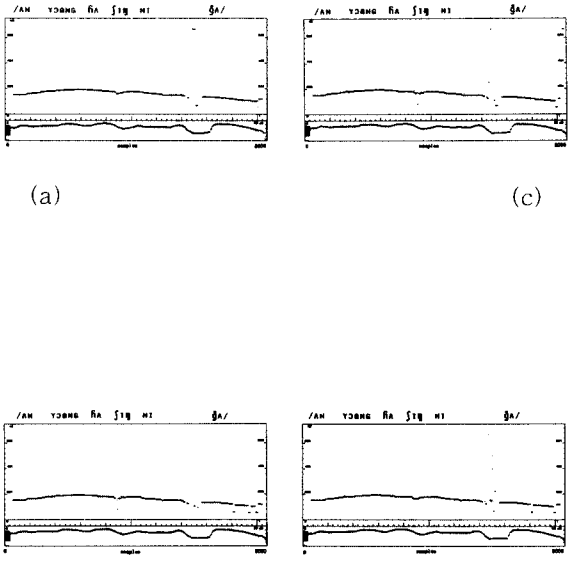
그림 4-6. 원 음성과 합성음들의 스펙트로그램
(a) 원음, (b) 32 레벨 VQ, (c) 64 레벨 VQ, (d) 32 레벨 FVQ

Fig. 4-6. Spectrogram of original speech
(a), speech synthesized by 32 level-VQ (b), 64 level-VQ (c), and 32 level-FVQ (d), respectivly

퍼지 VQ에 의한 합성음은 상대적으로 포만트 주파수의 윤곽이 뚜렷한 반면에 VQ에 의한 합성음은 포만트 주파수 사이에 음성 주파수 성분이 아닌 잡음 성분이 많이 분포하여 포만트 주파수의 윤곽이 뚜렷하지 않음을 볼 수 있어 퍼지 VQ에 의한 합성음이 원음의 자연성을 상대적으로 향상시킬 수 있음을 보여준다.

그림 4-7 (a), (b), (c), (d) 는 각각 원 음성, 32 레벨 VQ, 64 레벨 VQ, 그리고 32 레벨 FVQ 에 의한 합성음성의 피치 주파수를 나타낸다.

합성음의 피치 주파수는 합성방식에 큰 차이가 없이 원음의 피치 주파수(약 200 Hz) 를 유지하는 것을 볼 수 있다. 그러나 /안녕하십니까/ 의 문장 중에서 끝 부분의 발성부, 즉 /니까/ (5000샘플 - 6000샘플) 사이의 묵음부분에서 VQ 합성음에서는 음성정보가 나타나지 않은 반면에 (실제로는 600 Hz 이상의 고주파수로 나타남) 퍼지 VQ에 의한 합성음에는 음성정보가 포함되어 있음을 볼 수 있다.



(a) (c)

(b) (d)

그림 4-7. 원 음성과 합성음들의 피치주파수
Fig. 4-7. Pitch frequency of original speech
(a), speech synthesized by 32level-VQ (b), 64 level-VQ (c), and 32 level-FVQ (d), respectivly

이것은 FVQ에 의한 합성음의 피치 주파수가 퍼지판정의 smoothing 효과로 인해 분석 프레임 사이의 피치 주파수의 연속성을 유지하려는 것으로 VQ 합성음 보다 합성음의 열화를 방지함을 확인할 수 있다. 64 레벨 VQ에 의한 합성음성의 피치 주파수를 보면 알 수 있듯이 피치주파수의 변화를 원음에 가깝도록 하기 위해서는 코드북 크기를 늘이므로써 가능함을 볼 수 있다.

코드북 작성시 사용되는 알고리즘들 (K-means 알고리즘, 경쟁 학습신경망 알고리즘, FCM 알고리즘)에 의한 코드북의 성능비교는 참고문헌[19]을 참조 바란다. 코드북 자체의 성능은 거의 없음

을 알 수 있다.

5. 결 론

본 논문에서는 퍼지벡터양자화를 이용한 음성파형 합성방법을 제시하고, 고음질의 합성음을 얻기 위하여 퍼지벡터양자화의 Fuzziness 파라메타의 선정기준에 관하여 연구하였다. 퍼지벡터양자화를 이용한 음성합성은 VQ방식의 데이터 압축률을 유지하면서, 음성분석 프레임과 압축된 음성데이타(코드벡터)의 퍼지판정에기인한 합성 프레임사이의 smoothing 효과에 의해 프레임의 연결부에서 합성음의 열화를 감소시켜 음질을 높힐 수 있음을 보였고, 퍼지벡터양자화의 파라메타인 Fuzziness 값에 따라 합성음의 SQNR이 변화함을 실험으로 확인하여, SQNR을 최대로 하는 Fuzziness 값은 분석 프레임의 분산값이 크면 작게 선정하고, 크면 상대적으로 작게 선정하여야 함을 밝혔다.

향후연구는 지금까지의 연구결과를 토대로 분석 프레임의 분산값에 따라 적응적으로 최적 Fuzziness값을 설정하여 음성합성을 수행하는 적응 퍼지벡터양자화(Adaptive FVQ) 음성합성방식에 대한 연구와 데이터 압축률을 더욱 높히기 위해 음성파형의 샘플값이 아닌 음성의 특징 파라메타를 압축하여 퍼지 벡터양자화로 음성을 합성하는 적응 FVQ-CELP 방법에 관한 것이다.

참고문헌

[1] S. Furui, Digital speech processing, synthesis, and recognition, Marcel dekker, 1992.

[2] Juin-Hwey Chen and Allen Gersho, "Real- time vector APC speech coding at 4800 bps with adpaptive postfiltering,"pp. 51. 3.1 - 51. 3. 4, ICASSP 1987.

[3] N. S. Jayant and V. Ramamorthy, "Adaptive postfiltering of 16kbits/s ADPCM speech," pp. 16.4. 1 - 16. 4. 4, ICASSP 1986.

[4] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates,"Proc. Int'l Conference on Acostics, Speech, and Signal Processing, Tampa, March 1985.

[5] R. M. Gray, "Vector quantization, " IEEE ASSP Mag., vol. 1, pp. 4-29. April 1984.

[6] 이진이, 김형석, 이광형, " Fuzzy - C - means 알고리즘에 의한 벡터양자화 코드북의 성능비교, " 제 3회 인공지능, 신경망 및 퍼지시스템 종합 학술대회 논문집, vol 1. pp. 18-20, 1993, 10.

[7] 이진이, 김형석, 이광형, "신경망 학습 벡터양자화기에 의한 음성합성의 성능비교," 인공지능, 신경망 및 퍼지관련 학술발표회 논문집, vol. 1, pp. 10-18, 1993, 5.

[8] H. P. Tseng, M. J. Sabin, and E. A. Lee, " Fuzzy vector quantization applied to hidden Markov modeling," Proc. ICASSP 1987, Paper 15. 5.

[9] 이기영, "사상 멤버쉽 함수에 의한 화자적응 단어인식, " 명지대학교 박사학위 논문, 1991.

[10] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," IEEE, Tran. commun., vol. com-28, pp. 84-95, Jan. 1980.

[11] Bezdek, J. C., Pattern Recognition with Fuzzy Objective Function Algorithms , 1981, New York, London.

[12] Stanley C, Ahalt, Ashok K, Krishnamurthy, Prakoon Chen,and Douglas E.Melton, "Competitive Learning Algorithms for Vector Quantization," Neural Networks, Vol.3, pp. 277- 290, 1990.

[13] De Sieno. D.,"Adding a conscience to competitive learning. In IEEE Inter'l Conference on Neural Networks, pp. 1117-1124, 1988.

[14] H. J. Zimmermann, Fuzzy set theory and its applications, second edition, Kluwer academic publisers, 1991.

[15] Lee Jin-Yi, Lee Kwang-Hyung, " The Optimum Fuzzy Vector Quantizer for Speech Synthesis ," Fifth IFSA Congress, pp. 1321-1325, 1993.

[16] Ariel Corporation, User's Manual for PC-56D DSP Coprocessor Board, second edition November 1991.

[17] Motorora, DSP 56 KCC C cross compiler user's manual, 1989.

[18] Sensimetrics Coporation, Speech Station USER'S Guide, 1993.

[19] 이진이, "퍼지 벡터양자화와 신경망을 이용한 음성합성에 관한 연구," 숭실대학교 박사학위 논문, 1993.

이 진 이 (Jin-Yi Lee) 종신회원

1985년:숭실대학교 전자공학과 졸업
1988년:숭실대학교 전자공학과 석사
1994년:숭실대학교 전자공학과 박사
1995년~현재 충남산업대학교 전자공학과 전임강사
주관심 분야:퍼지신경망응용, 디지탈 통신, 데이타압축