다중대역 여기신호를 이용한 음성의 규칙합성에 관한 연구

o 경 연 정, 이 양 희
동덕여자대학교 전자계산학과

A Study on the speech synthesis-by-rule system using Multiband Excitation signal

Kyung youn-jeong, Lee yang-hee
Dongduck Women's Univ.

요 약

본 논문에서는 양질의 규칙합성음을 얻기 위하여, 유성음에 대한 여기신호로 임펄스 스펙트럼과 노이즈 스펙트럼을 다중대역으로 혼합하여 생성한 여기신호를 규칙합성에 적용하는 방법을 제안한다. 이 방법에서는, 분석합성에서 각 프레임별로 요구되었던 혼합여기신호에 대한 정보량 문제를 해결하기 위해 유성음의 정상부분의 한 프레임에 대해 혼합여기신호를 구하여 규칙합성에 적용하였고, 정보량을 더욱 줄이는 방안으로, 켑스트럼 유클리디안 거리를 이용하여 유성음을 분류하여, 각 그룹에 대한 대표 여기신호를 규칙합성의 여기신호로 사용하였다. 제안된 방법으로 음성을 합성한 결과 양질의 합성음을 얻을 수 있음을 확인하였다.

1. 서 론

언어음은 음향적인 에너지원으로서의 음원의 생성과 음원의 음에 음운성을 주는 조음의 2개 요소로 생성될 수 있다고 생각된다[1]. 그러므로 언어음 생성을, 음원과 조음부분으로 분리하여 모델화할 수 있다. 이처럼 음성발생의 일반적인 모델은 음원과 성도를 독립적인 것으로 가정하여 각각을 모델화 하는 것으로, 조음특성에 해당하는 성도는 합성 디지탈 필터에 의해서 모델링되고, 음원특성에 해당되는 음원발생 모델은 유성음인 경우 주기적 임펄스열을, 무성음인 경우 백색잡음을 사용한다[2].

여기신호로 백색잡음과 임펄스만을 사용하는 단순 음원에 의해 생성되는 합성음은 음질면에서 한계성이 있게 된다. 즉, 일반 음성생성 모델에 의한 합성음은 어느 정도 명료한 합성음은 가능하나 단순한 음원의 사용으로 양질의 음성을 합성하기는 어렵다. 특히, "buzziness"등이 문제시 된다[3].

본 논문에서는 규칙합성에서 필수적인 피치변화에 의한 왜곡을 최소화할 수 있고 유성 프레임이라도 스펙트럼 상에 노이즈 성분을 포함하고 있음을 고려하여, 분석합성에서 제안된[3][4] 각 유성음의 프레임별로 구해진 혼합여기신호를 양질의 음성을 합성하기 위해 규칙합성시스템에 적용하는 방법을 제안한다. 이 방법에서는, 혼합여기신호에 대한 정보량을 최소화하기 위하여 유성음을 분류하고 각 그룹의 대표 여기신호를 선택, 규칙합성시 적용하여 규칙합성음의 음질을 개선하고자 한다.

2. 혼합여기신호를 이용한 유성음의 음질개선

일반적으로 유성에 대한 음원은 임펄스를 사용하고 있다. 그러나 실제 음성을 분석한 그림 1에서 보면 유성에 해당되는 프레임의 스펙트럼이라도 그 안에 하모닉스(harmonics) 뿐 아니라 노이즈 성분도 갖고 있음을 알 수 있다. 그림 1의 (a)는 25.6 msec의 블랙크만 창함수가 적용된 모음의 한 프레임에 해당되는 파형이고 (b)는 그의 스펙트럼을 보이고 있다. 이처럼 유성 프레임의 스펙트럼이라도 그 안에 노이즈 스펙트럼에 해당되는 성분을 포함하고 있다. 그러므로 이러한 유성에 대해 음원으로 임펄스만을 사용한다면, 이 프레임 내의 노이즈 성분이 무시되어 고품질의 합성음을 기대하기 어렵다.

분석합성에서 제안된 방법에 따르면[3][4], 유성음의 한 스펙트럼에서 유/무성 구간을 판별한 뒤, 각 구간에 대해 임펄스에서 얻은 스펙트럼과 백색잡음에서 얻은 스펙트럼을 혼합한다. 즉, ω를 주파수 대역의 각주파수로 놓고, 원음의 스펙트럼을 $|Sp[\omega]|$라 하면 이 안에는 임펄스와 노이즈 성분이 모두 포함되어 있다. 이 때 합성에 의해 얻어지는 스펙트럼을 원음에 비교하여 $|Sp'[\omega]|$로 나타내고 스펙트럼 인벨로프를 $|Se[\omega]|$로, 임펄스의 스펙트럼을 $|Ip[\omega]|$, 백색잡음의 스펙트럼을 $|Np[\omega]|$로 나타낼 때, 지금까지는 그 프레임이 유성이라 판별되면 $|Sp'[\omega]| =$

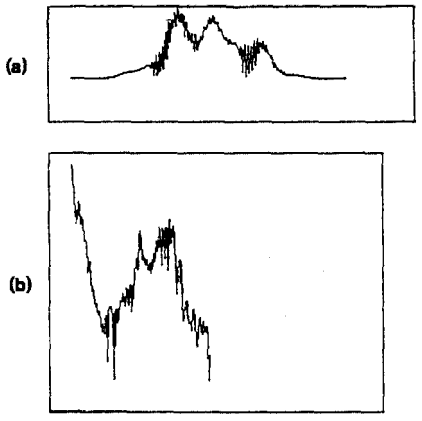


그림1 유성음의 스펙트럼 ((a)파형,(b)스펙트럼)

|Ip[ω]| * |Se[ω]|로 얻었으나 본 논문에서는 프레임 내 유/무성 구간 표시 함수인 VI[ω], VN[ω]을 적용하여 합성음의 스펙트럼을 다음 식에 의해 얻는다.

$$|Sp'[\omega]| = [|Ip[\omega]|*VI[\omega] + |Np[\omega]|*VN[\omega]] * |Se[\omega]| \quad (1)$$

이 때, VI[ω]와 VN[ω]은

$$VI[\omega] = \begin{cases} 1 & : \text{유성} \\ 0 & : \text{무성} \end{cases} \quad (2)$$

$$VN[\omega] = \begin{cases} 0 & : \text{유성} \\ 1 & : \text{무성} \end{cases} \quad (3)$$

으로 정의된다.

이렇게 혼합된 스펙트럼을 역 푸리에 변환하면 시간축 상의 신호를 얻을 수 있고, 이를 음원으로 하여 합성을 한다.

혼합여기신호를 얻는 과정은 다음과 같이 나타낼 수 있다.

$$|sp'[\omega]| = |Ip[\omega]|*VI[\omega] + |Np[\omega]|*VN[\omega] \quad (4)$$

$$W[t] = IFFT[|sp'[\omega]|] \quad (5)$$

이 때, 시간축 상의 결과 W[t]는 win[t]*x[t] 이므로 원래의 신호 x[t]는 x[t] = W[t]/win[t] 로 얻을 수 있다. 이는 원음의 음원에 가장 가까운 신호를 얻기 위한 것이다.

합성시 사용음원은 펫치길이 만큼이므로 평균펫치 길이의 창함수 winpit[]를 x[]와 곱하면, Vs[t] = winpit[t]*x[t]로써 음원으로 사용될 시간 축상의 신호 Vs[]를 얻을 수 있다. 최종적으로 얻어진 Vs[t]를 유성 프레임에 대한 음원으로 사용하게 된다. 수식에서 알 수 있듯이 Vs[t]는 임펄스와 노이즈 스펙트럼이 혼합된 스펙트럼 |sp'[ω]|로 부터 역 푸리에 변화에 의해 얻어진 신호로 이를 다시 분석하여 스펙트럼을 구하면 그림 2의 (a)와 같이 나타난다. 이는 최종적으로 얻은 신호 Vs[t]가 어느 특정 주파수 대역에 치우치지 않는 평탄한 스펙트럼을 갖고 있으므로 음원으로 사용가능함을 보여준다. (b)는 실제 합성시 사용되는 혼합여기신호의 스펙트럼으로 임펄스와 노이즈에 대한 스펙트럼이 혼합되어 있음을 볼 수 있다. (c)는 유성에 대한 잔차신호의 파형과 해당 스펙트럼이다. Vs[t]를 음원으로 하는 합성음은 처음 정의한 식의 |Sp'[ω]|를 스펙트럼으로 갖게된다.

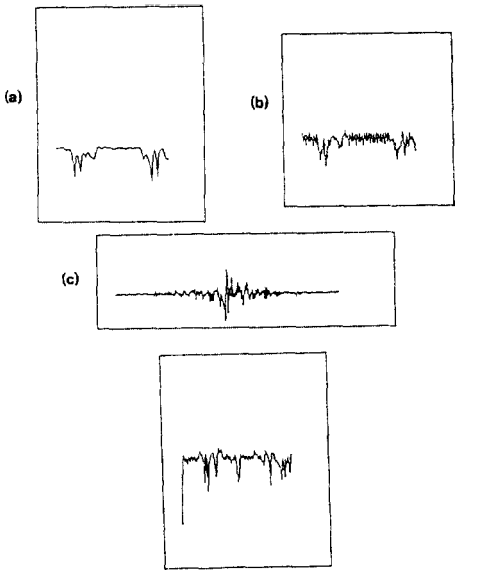


그림2 혼합여기신호의 스펙트럼 ((a)혼합여기신호의 스펙트럼, (b)실제 적용시 사용되는 혼합여기신호의 스펙트럼, (c)유성에 대한 잔차신호의 파형과 스펙트럼)

그림 3에 이러한 유성음에 대한 음원을 얻는 과정을 보이고 있다. (a)에는 원음의 스펙트럼, (b)는 이에 대한 유/무성 정보, (c)와 (d)는 각각 임펄스와 백색잡음에 대한 스펙트럼, (e)는 유/무성 구간 정보에 의해 둘을 혼합한 스펙트럼이다. 이 때 얻은 스펙트럼을 역 푸리에 변환하여 다시 시간축 상의 신호로 변환하여 최종적으로 음원으로써 사용하는 혼합여기신호는 그림 4의 파형이다.

위 과정에 의해 얻어진 혼합여기신호를 규칙합성에 적용한 결과 양질의 합성음이 생성됨을 확인하였다.

3. 대표 여기신호를 위한 유성음 분류

본 논문에서는 위와 같은 방법으로 얻어진 유성에 대한 혼합여기원을 규칙합성 시스템에서 사용하기 위하여, 각 유성음별로 그에 해당하는 음원정보를 갖는 것 보다,

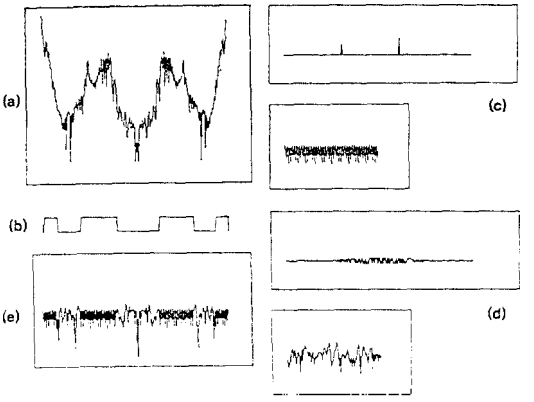


그림3 혼합여기신호를 얻는 과정 ((a)원음의 스펙트럼, (b)유/무성 정보, (c)임펄스 스펙트럼, (d)노이즈 스펙트럼, (e)혼합 스펙트럼)

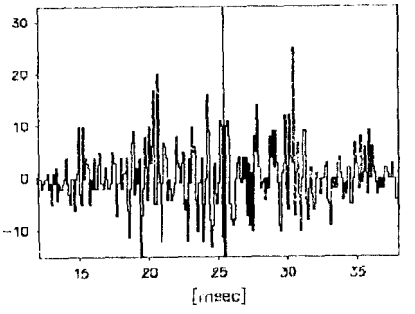


그림4 혼합여기신호

정보량을 줄이기 위하여 유사한 모음들을 묶어 같은 그룹내의 유성음에 대해서는 한 음원을 사용하는 방법을 제안한다. 모음별의 유사도는 정상부근의 스펙트럼 인벨로프의 켑스트럼 값의 유클리디안 거리를 구해 그를 기준으로 하고, 비교적 그 차이가 적은 모음들을 같은 그룹으로 분류하였다. 켑스트럼 계수의 켑스트럼 유클리디안 거리는 두 스펙트럼 사이의 왜곡 정도를 잘 나타내고 있는 것으로 알려져있다[5]. 본 논문에서 사용한 켑스트럼 거리는 식 (6)에 보이고 있다.

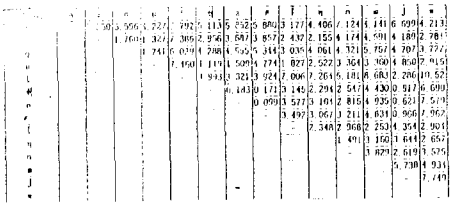
$$Ed = 10/\ln 10\sqrt{2\sum_{i=1}^{p}(c^{v1}[i] - c^{v2}[i])^2} \qquad (6)$$

이 때, v1 ∈ {모든 유성음}, v2 ∈ {v1을 제외한 유성음}이 된다.

켑스트럼 거리척도를 이용하여 나눈 그룹 중 같은 그룹의 모음들은 한 음소에 대한 혼합여기원을 음원으로 하여도 그 음질에 큰 영향이 없는 것으로 실험결과 나타났다. 표 1에 식 6을 이용하여 각 유성음간 유클리디안 거리를 구한 것을 나타내었다.

이처럼 유성음 간의 분류가 가능한 이유는, 그림 5에서와 같이 동일 모음이라도 다른 환경하에서는 스펙트럼 인벨로프 사이의 차이가 나타나 두 스펙트럼의 켑스트럼

표 1 유성음간 켑스트럼 유클리디안 거리



계수 사이에 크지는 않지만 유클리디안 거리값이 있게 되는 등 적은 값의 켑스트럼 거리를 갖는 음소들은 유사한 것으로 분류할 수 있기 때문이다. 그림 5의 (a)는 /아/음의 단독 발성음의 스펙트럼 인벨로프를 나타낸 것이고 (b)는 /가/음의 스펙트럼 인벨로프를 나타내고 있다. 같은 모음 /아/이지만 앞에 자음이 있음으로 인해 스펙트럼 인벨로프가 차이가 있고 이러한 현상은 모음의 천이부와 정상부, 끝부분에도 나타나게 된다. 이처럼 동일모음 간에 나타나는 유클리디안 거리값을 기준으로 하여 유성음을 분류하였다. 유성음 분류시의 문턱치는 대략 2.5db정도를 적용하였고, 표 2에 유성음 분류 결과를 나타내었다. 이 때 적용하는 문턱치는 유성음 분류를 어느 정도 세분화할 것인가에 따라 달라질 수 있다. 본 실험에서 행한 유클리디안 거리를 구하는 실행 예는 그림 6에 보이고 있다.

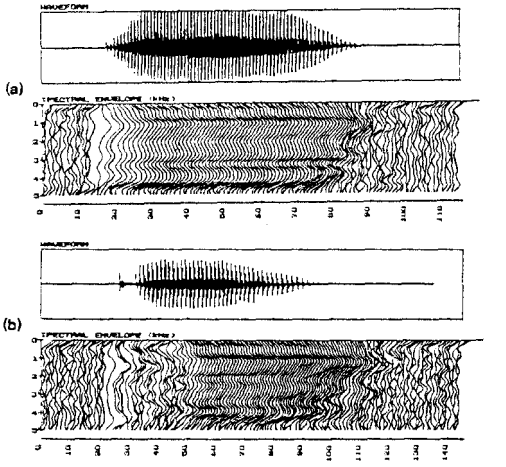


그림5 동일모음의 다른 환경에서의 스펙트럼 인벨로프 ((a)/아/, (b)/가/)

표 2 유성음 분류

그룹 1	j, ɛ, ɐ, e
그룹 2	u, ɯ, o
그룹 3	l, n, w
그룹 4	n
그룹 5	m
그룹 6	a
그룹 7	i

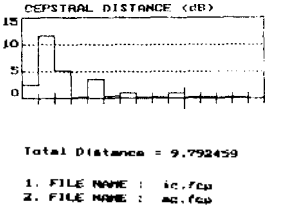


그림6 켑스트럼 유클리디안 거리를 구하는 예

유성음의 음소별로 혼합여기신호원을 달리하여 합성한 결과와 그룹내의 한 음소에 대한 혼합여기신호를 사용하여 합성한 합성음을 비교해 본 결과 별로 차이가 없는 것으로 나타났다.

4. 실험 및 결과

유성음 분류에 의한 각 그룹의 대표 혼합여기신호를 여기신호로 실제 이용하여 규칙합성을 행한, 합성파형은 그림 7에 보이고 있다. 그림 7의 (a)는 원음, (b)는 임펄스 구동에 의한 분석합성음, (c)는 본 논문에 의한 혼합여기원에 의한 합성파형이다. 또한 혼합여기신호를 사용한 경우 피치변화에 따른 음질저하도 줄일 수 있다. 그림 8은 고유피치에 1.5배하여 합성한 합성음을 나타내고 있는데 (a)의 임펄스 구동음인 경우 피치주기가 길어짐에 따라 그 합성음의 음질저하가 심해지나 (b)의 혼합여기신호에 의한 합성음은 피치 변화에 대해서도 양질의 합성음을 생성함을 알 수 있다. 이는 규칙합성시 나타나는 피치변화에 따른 합성음의 스펙트럴 왜곡이 혼합여기신호 사용으로 해결가능함을 보이고 있다. 그림 9는 혼합여기신호를 유성음의 음원으로 하여 규칙합성을 한 예이다. (a)는 임펄스 구동에 의한 규칙합성음, (b)는 혼합여기신호에 의한 규칙합성음이다. 이는 규칙합성음인 /안녕하십니까/의 일부분 /안녕하십/에 해당되는 파형을 보이고 있다.

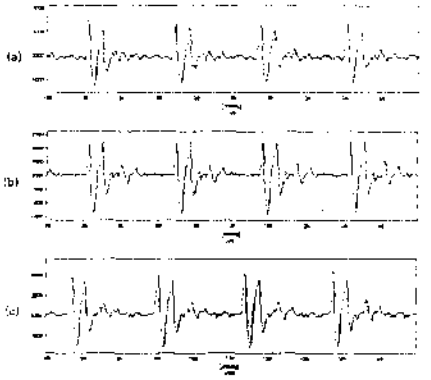


그림7 혼합여기신호에 의한 합성음 ((a)원음, (b)임펄스에 의한 합성음, (c)혼합여기신호에 의한 합성음)

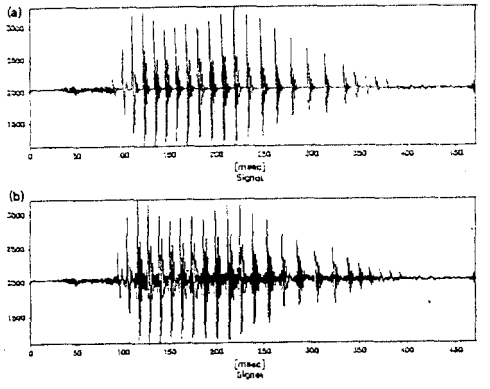


그림8 피치변경시 혼합여기신호에 의한 합성음 ((a)임펄스에 의한 합성음, (b)혼합여기신호에 의한 합성음)

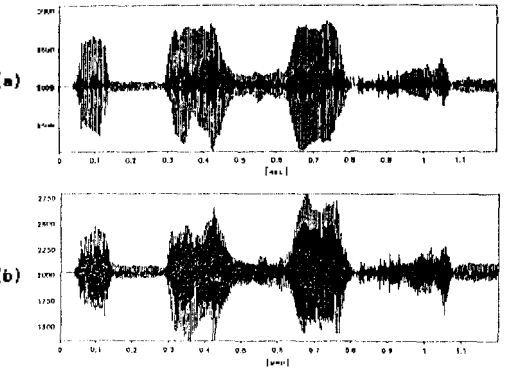


그림9 혼합여기신호를 규칙합성에 적용한 예 ((a)임펄스에 의한 합성음, (b)혼합여기신호에 의한 합성음)

합성파형과 주관적 청취실험을 통해 단순한 임펄스 음원을 사용한 합성음보다 이처럼 혼합여기원을 사용할 경우 훨씬 양질의 합성음을 얻을 수 있음을 알 수 있었다.

5. 결 론

본 논문에서는 규칙합성시스템 구현에서 발생되는 음질 향상에 대한 문제점을 살펴보고 그 해결방안을 제안하였다. 제안된 방법에서는 유성음 합성의 경우, 유성 프레임이라도 노이즈 성분을 갖고 있으므로 분석합성에서 제안된 임펄스와 노이즈의 스펙트럼이 혼합된 혼합여기원을 사용하여 단일 임펄스 사용시 나타나는 buzziness의 해결과 함께 합성음의 음질을 개선시켰다. 이는 규칙합성시에 필수 불가결한 피치변경으로 인한 스펙트럴 왜곡도 줄일 수 있는 것으로 나타났다. 혼합여기원에 대한 추가 기억용량에 관한 고려에서 캡스트럼의 유클리디안 거리에 의한 유성음을 분류하여 동일 그룹의 유성에 해당하는 혼합여기원을 여러 유성음에 대해 적용하여도 그 합성음의 품질이 매우 우수한 것으로 나타났다.

앞으로 유성음 프레임 내의 자동 유/무성 판별이 이루어지면 더욱 우수한 혼합여기신호를 얻을 수 있을 것으로 기대된다.

[참고문헌]

[1] 이양희외, Man-Machine Interface를 위한 음성처리연구, 한국전자통신연구소 보고서, pp.311-320, 1992.

[2] L. R. Rabiner, R. W. Schafer, "Digital processing of speech signal", Prentice-Hall, 1978.

[3] Denial W. Griffin, et al., "A High Quality 9.6 kbps Speech Coding System", IEEE, pp.125-128, 1986.

[4] Denial W. Griffin, et al.. "A New Model-Based Speech Analysis/Synthesis System", IEEE, pp.513 - 516, 1985.

[5] S. Furui, "Digital Speech Processing, Synthesis, and Recognition", MARCEL DEKKER, pp.168-169, 1992.