기존 가수 및 신세대 가수의 소리지속시간 분석에 관한 연구

정영훈, 배명진

숭실대학교 정보통신공학과

A Study on Analysis of Speech Duration Between the Existing Singer and New Generation Singer

YoungHoon Jung, MyungJin BAE

Dept. of Information & Telecommunication Engineering, Soongsil Univ., Korea

mjbae@saint.ssu.ac.kr

요 약

음악을 함에 있어서 정확하고 매력적인 발성을 하는 것도 중요하지만 더욱 기본적이고 중요한 것이 정확한 발음을 내는 것이다. 정확한 발음이 해결되지 않은 상태에서는 아무리 발성법을 꿰뚫고 있다하더라도 많은 사람들에게 자신이 전달하고자 하는 메시지를 제대로 전달하지 못하게 된다. 보통 노래를 잘 부르기 위해서 노래방 같은 곳을 찾아가 노래 연습을 하는 사람들이 많이 있는데, 무엇보다 기본적인 발음이 명확하지 않으면 노래를 잘 부른다고 볼 수는 없는 것이다. 랩을 주로 하는 신세대 가수들의 음악을 들어 보면 자막을 보지 않고서는 무슨 말인지 알아들을 수가 없다. 그들이 노래할 때 입 크기의 변화 없이 입술모양만 변화시키면서 발성하기 때문이다. 음성은 기본적으로 여기성분과 성도성분으로 구분할 수 있다. 성도는 인두강과 구강을 합쳐서 일컫는다. 따라서 입 모양을 어떻게 하느냐에 따라서 같은 말이라도 명료성이 달라지게 된다. 본 논문에서는 이 소리지속시간을 비교 평가하기 위해서 기존가수와 신세대 가수의 한 음절에 대한 지속시간을 비교하여 보았고 8Khz까지의 스펙트로그램을 비교하였다. 비교결과 기존 가수가 신세대 가수에 비하여 말의 의사 전달에 있어서 명료하게 전달 할 수 있다는 것을 알 수 있었다.

1. 서 론

일반적으로 자연스러운 대화를 할 때나 글을 읽을 때의 음성에는 피치, 에너지, 지속시간 등의 운율정보가 포함되어 있다. 현재 대부분의 신세대 가수들의 음악은 노래 노래만 듣고 있으면 무슨 말을 하는지 거의 의미파악이 되지 않는다. 음악을 들으면서 자막을 봐야만 비로소 무슨 말을 하는지 알 수가 있다. 하지만 기존의 가수나 발라드 가수의 노래의 경우 자막 없이도 노랫말을 알아들을 수 있다. 이는 화자의 의미 전달력 즉 명료성에 관련이 있는 부분이다. 음성은 여기 성분과 성도성분으로 구분할 수가 있다. 성도의 성분은 인두강과 구강으로 구분되어 지는데 입 모양을 어떻게 하느냐에 따라서 성도의 특성이 달라지고 포만트의 공명특성 또한 달라지게 되어 같은 발성이라도 음성학적 정보를 파악하는 정도가 달라지게 된다. 본 논문에서는 기존 가수 한 명과 신세대 가수 3명의 무반주 음성에 대하여 한 음절의 지속시간과 8khz까지의 스펙트로그램을 비교 평가하였다.

2. 음성 생성 모델

2-1. 음성생성 시스템의 해부학적 측면

음성 발생 시스템을 음향필터 작동으로 기술하면 매우 유용하다. 이에 대한 간단한 개념도를 그림 2-1에 나타내었다. 세 개의 주요 강(인두강, 구강, 비강)은 세 개의 주요 음향 필터를 구성한다. 필터는 그 아래 기관에서 여기 되고 출력단은 입술이 부하로 연결된다. 조음기관은 대부분 필터와 직접 연관이 있고 시간에 따른 시스템의 특성, 여기 형태, 출력단 부하 특성들을 바꾸는데 관여한다.

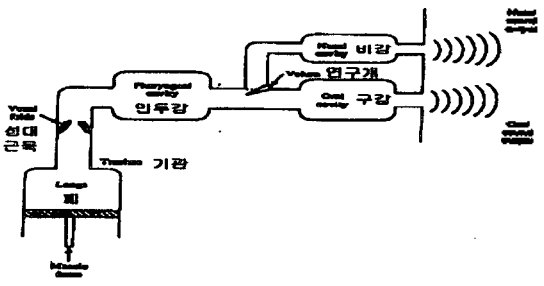

그림 2-1. 인간 음성 발생의 블록도

위와 같은 발성시스템을 기초로 해서 음성 신호는 크게 유성음(Voice Sounds)과 무성음(Unvoice Sounds)으로 구분할 수 있다. 유성음은 허파에서 압축된 공기가 성대를 통과하면서 공진이 발생하고 이는 곧 성대를 주기적으로 떨게 함으로써 발생하는 소리를 말한다.
모음"아"를 발음할 때 목 부분이 떨리는 것을 볼 수 있는데 이런 음성을 유성음이라고 한다. 무성음은 공진이 발생하지 않을 정도의 빠른 속도로 공기를 압축하고 성도의 일부를 좁히면서 또 한번 압축해서 난기류를 만들어 내는 소리를 말한다. "흐"와 같은 음이 대표적이며 성대의 떨림도 없고 유성음에 비해 많은 공기가 입으로 나오는 것을 알 수 있다.
이런 과정을 거쳐 발생한 음성신호는 어떤 모양을 하고 있는지를 시간 축에서 나타낸 것이 [그림 2-2]이다.

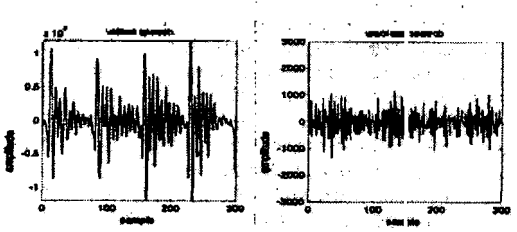

그림 2-2. 무성음과 유성음의 신호파형

2-2. 음성신호의 발생 모델

음성은 여기(excitation)에 따라 세 가지로 나눌 수 있다. 첫째, 유성음은 조정된 성대의 팽창과 함께 성문을 통한 공기의 힘에 의해 생성되어 갑쇠 진동하여 성도를 자극하는 공기의준 주기적인 펄스를 만든다. 유성음은 /아/, /에/, /이/, /오/, /우/ 등의 모음과 /ㄹ/, /ㅁ/등의 비음으로 구성된다. 둘째, 마찰음 또는 무성음은 성도에 있는 어떤 점에서 협착을 형성하고, 통요를 만들기 위해 고속으로 협착점을 통과하는 공기의 힘에 의해 발생된다. 이것은 성도를 자극하기 위해 광대역 잡음원을 생성하게 된다. 셋째, 파열음(plosive)은 완전히 입을 폐쇄하고, 이 폐쇄 뒤에서 압력을 마들어 갑자기 느슨하게 함으로써 생성된다.
음성생성에 있어서 성도의 공명주파수를 포만트 주파수 또는 포만트 라고 한다. 포만트 주파수는 성도의 모양과 면적에 따라 다르고, 소리의 형태는 성도의 모양이 변화함으로써 시간에 따라 변한다.

2-2-1. FORMANT

Formant는 성대의 기하학적인 모양에 따라 달라지고 특정 음성신호는 대표적인 몇 개의 Formant로 대표되어질 수 있다. 예를 들어 "아"라는 음과 "어"라는 음은 사람의 성도 변화에 의해서 만들어 낼 수 있으며 이 때의 Formant 주파수는 각각 다른 양상을 나타낸다. 따라서 Formant는 음성신호 모델링에서 중요한 요소로 작용한다. [그림2-3]은 일반적인 음성 신호의 모양과 주파수 스펙트로그램을 보여준다. (a),(b)그림은 유성음, 무성음의 모양을,(c),(d)그림은 가가 신호의 주파수 스펙트로그램을 나타낸다. (c)그림을 자세히 보면 스펙트로그램의 전체적인 모양이3개의 봉우리를 가짐을 알 수 있다.
약 700Hz, 1200Hz, 2600Hz에 각각 하나의 큰 봉우리가 보이는데 이것이 바로 주파수 영역에서 관찰한 Formant 주파수이다. 무성음인 (b)그림은 Formant 정보의 보이지 않고 주파수 영역에서 역시 노이즈와 비슷한 모양을 나타낸다.[10]

2-2-2. Pitch

일반적으로 음성신호의 Pitch라는 단어는 Fundamental 주파수라는 말과 동의어로 쓰인다.
Fundamental 주파수는 음성신호 중에서 가장 기본이 되는 주파수, 즉 시간 축에서 커다랗게 나타나는 peak들의 주파수를 의미하며 이미 설명한 바 있는 성대의 주기적인 떨림에 의해서 생성된다. Pitch는 인간의 청각에 매우 민감하게 반응하는 파라미터로써, 음성신호의 화자를 구분하는데 사용하며, 음성신호의 naturalness에 큰 영향을 미친다. 그러므로 정확한 Pitch 해석은 음성합성의 음질을 좌우하는 중요한 요소이며 음성코딩에 있어서도 Pitch의 정확한 추출과 복원은 음질에 결정적인 역할을 한다. 그리고 Pitch 정보는 음성신호의 유성음/무성음을 판단하는 파라미터로도 사용된다. Pitch는 허파에서 압축되어진 공기가 성대에 진동을 일으키면서 생기는 주기적인 Pulse이므로 성대의 진동 없이 난류를 일으키는 무성음의 경우에는 Pitch가 생기지 않는 것은 당연하다 하겠다. Pitch는 사람의 성대 구조상 일정한 제약을 가지는데, 남성의 경우 일반적으로 50 — 250Hz, 여성의 경우 120 - 500Hz에 존재한다. 그리고 Pitch는 강세, 억양, 감정등에 따라서 변한다.

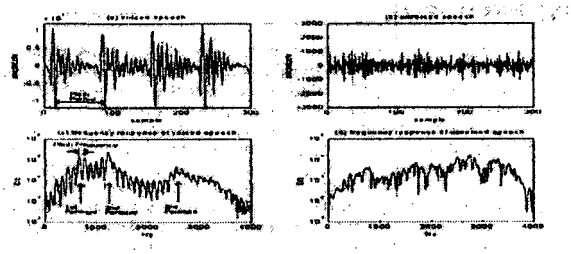

그림2-3. 유성음, 무성음의 주파수 영역 모양과 Formant, Pitch와의 관계

2-2-3. 음절의 지속시간

발음된 문장 속에서의 각 음절 및 음소의 지속시간은 화자의 심리상태, 문장의 의미와 같은 주관적인 요소와 발음된 음절의 수, 음절의 문장 내 위치, 발음속도, 문장의 문법적 구조 그리고 액센트의 유무 등과 같은 객관적인 요소들에 의하여 변화하게 된다. 예를 들어서, 음절의 수가 많거나 발음속도가 빠를수록 각 음소의 지속시간은 각 음소의 지속시간은 짧아지게 되며, 문장의 끝 음절이나 액센트를 받는 음절의 지속시간은 상대적으로 길어지게 된다. 문장을 구성하는 각각의 음절 및 음소의 지속시간은 각기 서로 다른 지속시간을 갖게 되는데, 음절의 지속시간은 대체로 액센트가 있는 음절은 액센트가 없는 음절에 비해 길어지며 음절의 지속시간은 말토막 내에서 음절의 개수가 많을수록 짧아진다. 초, 중, 종성의 종류에 따라 음절의 지속시간이 달라진다.
폐쇄음이 종성으로 오는 경우 지속시간이 짧아지며

유음이나 비음이 종성인 경우 길어진다. 종결형 어미나 연결형 어미 등의 끝에 위치하는 음절의 지속시간은 다른 음절에 비해 상대적으로 길어진다.[11]

2-2-4. 음성의 에너지와 명료도

에너지는 사람의 음성발성에 있어서 의사전달을 정확하게 전달하는데 큰 역할을 한다. 사람의 회화영역은 100~8000Hz이나, 일반적으로 많이 사용되는 주요 회화영역은 300~3000Hz이다. 전체 영역중 500~4000Hz 부분이 중요하며, 어음명료도의 83%를 차지하게 된다. 에너지는 작지만 고주파 부분이 대화음 이해에 중요함을 알 수 있다. 정상적인 귀는 500~4000Hz부분이 가장 민감하게 작용한다.

3. 실험 및 결과

본 논문을 실험하기 위해서 현재 활동하고 있는 기존가수 한명과 신세대 가수 3명의 노래를 무반주로 채취하였다. 이를 비교하기 위해서 삼보 PC PentiumIV 1.7G를 사용하였고, 16bit AD 변환기를 사용하였다.

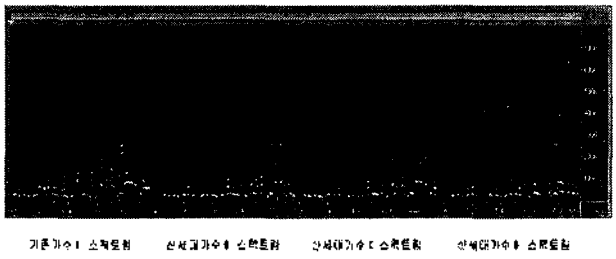

그림 3-1. 기존가수와 신세대가수의 노래 스펙트로그램

그림3-1은 4명의 가수가 부른 노래에 대한 8khz까지의 스펙트로그램이다. 4khz밑으로는 저주파를 보여주고 있다. 4khz 밑으로 보게 되면 왼쪽부분에 연하게 보이는 부분이 있는데 그 부분이 선명하게 보일 수 록 노래를 잘 소화 한다고 볼 수 있다. 그것은 낮은음이나 높은음에서도 에너지가 일정하다는 것을 의미한다. 4Khz이상 부분은 고주파를 보여주고 있는데, 첫 번째 스펙트로그램을 보면 가운데 부분이 검게 나타나질 않고, 왼쪽 밑을 보아도 색 변환이 거의 없음을 볼 수 있다. 4Khz이상 부분을 보아도 끊임이 없고 일정하다는 것을 알 수 있다. 그러나 나머지 신세대가수들을 보면 검게 일어나는 부분이 많으며 끊임 부분도 볼 수 있다. 스펙트로그램으로 보았을 때는 기존가수의 경우 전 대역에서 저주파 대역부터 고주파 대역까지 골고루 분포하였다. 신세대 가수의 경우는 고주파 대역에서는 나타나지 않는 부분이 많이 존재하였다.

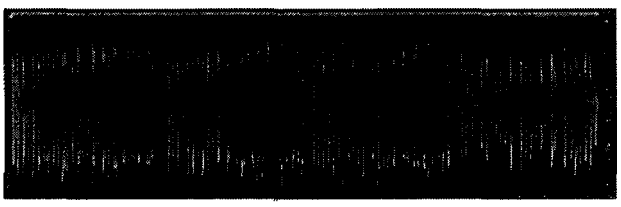

그림3-2. 기존가수와 신세대가수 시간 축 에너지 레벨

그림3-2는 시간 축에서의 에너지 레벨을 보여주고 있다. 시간 축에서 전체 에너지 레벨을 보면 기존가수는 다른 3명의 신세대 가수보다 에너지 레벨이 일정하다는 것을 볼 수 있다. 3명의 신세대 가수는 에너지 레벨이 고르게 나타나 있지 않다. 이것은 기존가수는 신세대가수보다 에너지가 일정하며 공명도 높다고 볼 수 있다. 그림3-3은 마지막부분의 "사랑해", 그림3-4는 중간부분의 "우리", 그림3-5는 중간부분의 "영원을"이라는 노랫말이다. 4명의 가수들의 발생 지속 시간이다.

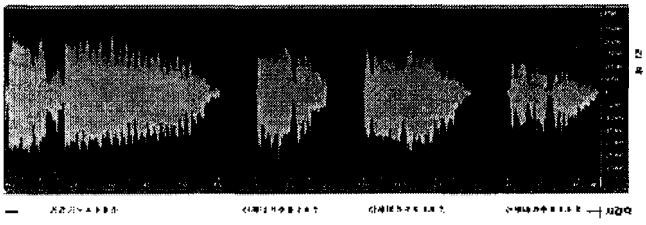

그림 3-3. 마지막부분의 "사랑해" 발생 지속시간

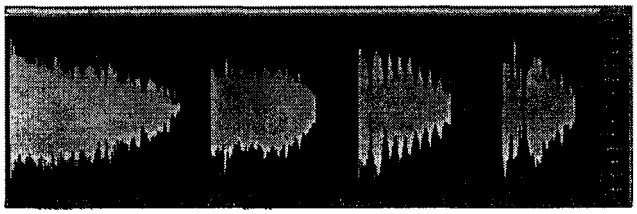

그림 3-4. 중간부분의 "우리" 발생 지속시간

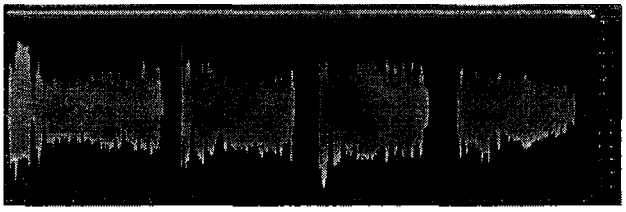

그림 3-5. 중간부분의 "영원을" 발생 지속시간

기존가수는 그림3-3은 지속적인 음이 6초 동안 지속적인 것을 볼 수 있으며, 떨림(진폭)도 크고 일정하다는 것을 알 수 있다. 신세대 가수는 각각 3초에서 4초 동안 지속되는 것을 볼 수 있을 것이다. 그림3-4, 그림3-5는 기존가수의 음이 3.5초 동안 지속되는 것을 볼 수 있다. 신세대가수는 2초에서 2,5초 동안 지속되는 것을 볼 수 있다. 기존 가수는 입 모양을 크게 하고 발성하며 고음에서 지속적인 음이 나머지 신세대 가수 3명보다 2초에서 3초가 더 길다. 기존 가수를 제외한 나머지 신세대가수 3명은 입 크기의 변화 없이 입술모양만 변화 시키면서 발성하였다. 그림3-6은 마지막부분의 "사랑해", 그림3-7은 중간부분의 "우리", 그림3-8은 중간부분의 "영원을"이라는 노랫말의 스펙트로그램이다.

기존가수 스펙트로그램을 보게 되면 에너지가 고르게 분포가 되어 있다. 신세대 가수들의 스펙트로그램을 보게 되면 저주파에서는 고르게 분포가 되어 있지만 고주파에서는 일정하지 않다는 것을 볼 수 있다.

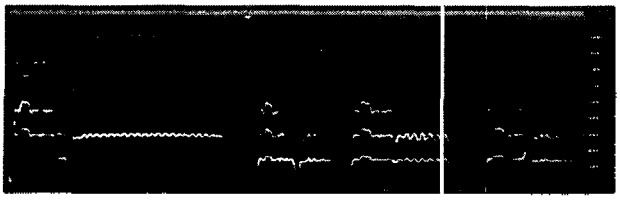

그림 3-6. 마지막 부분의"사랑해" 스펙트로그램

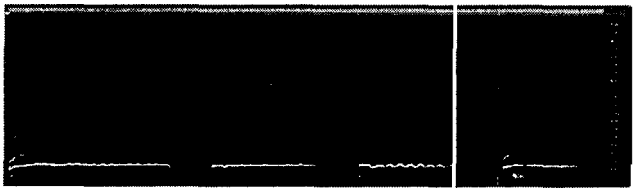

그림3-7. 중간부분의 "우리" 스펙트로그램

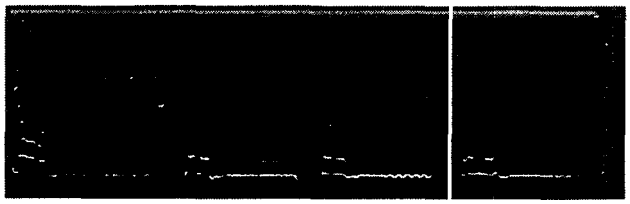

그림3-8. 중간부분의 "영원을" 스펙트로그램

그림3-9와 그림3-10은 노랫말 "사랑해"부분의 스펙트로그램 성분을 분석 한 것이다. 기존가수는 신세대 가수보다 peak가 일정하다는 것을 볼 수 있다. 신세대가수는 처음부분에는 peak가 보이지만 시간이 자나면서는 일정하지 않으면 거의 사라지는 것을 볼 수 있다. 기존가수를 보면 pitch가 일정하게 나타나지만 신세대가수는 기존가수보다 일정하게 나타나질 않는다.

전체 노랫말의 지속시간은 기존가수와 신세대가수의 차이는 없다. 전체 노랫말이 아닌 일부 노랫말에서는 기존 가수가 신세대 가수보다 지속시간이 약 2초에서 3초 정도가 더 지속 된다는 것을 볼 수 있다.

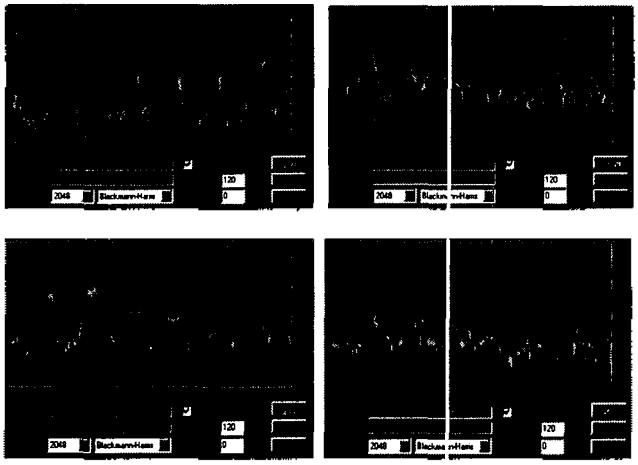

그림3-9. 기존가수와 신세대가수(1)(2)(3)의 스펙트로그램 성분 부석

4. 결 론

본 논문에서는 명료성과 지속시간을 비교 평가하였다. 기존가수와 신세대 가수의 한 음절에 대한 지속시간을 비교하였다. 노래에서 전체 지속시간은 기존가수와 신세대 가수가 비슷하다. 그러나 전체 노랫말이 아닌 일부 노랫말에서는 기존 가수가 신세대 가수보다 지속시간이 약 2초에서 3초정도가 더 지속된다는 것을 볼 수 있다. 스펙트럼으로 보았을 때는 기존가수의 경우 전 대역에서 저주파 대역부터 고주파 대역까지 골고루 분포한 반면 신세대 가수의 경우는 저주파에서는 골고루 분포하였지만 고주파 대역에서는 나타나지 않는 부분이 많이 존재하였다. 전체 노랫말의 스펙트럼에서의 에너지 레벨을 보면 기존가수는 에너지 레벨이 다른 신세대 가수들보다 높다는 것을 볼 수 있다. 이것은 기존가수는 공명주파수가 저주파에서 고주파까지 나타나며 Pitch가 일정하며 특징이 잘 나타나 있음 말해준다. 신세대 가수들은 공명주파수가 저주파에서만 나타나고 고주파에서는 거의 나타나지 않는다. Pitch가 일정하지 않고 특징이 없다. 저주파는 말의 명료성을 나타내는 부분이다. 따라서 기존가수가 신세대 가수에 비하여 말의 의사 전달에 있어서 명료하게 전달 할 수 있다. 기존 가수는 입 모양을 크게 하고 발성하며 고음에서 지속적인 음을 내며 폐에서부터 성도를 지나서 떨려오는 음으로 노래를 부름을 알 수 있다. 신세대가수는 입을 작게 벌리고 노래를 했으며 성대의 떨림이 아닌 맑은 소리로만 노래를 불렀음을 알 수 있다. 신세대가수는 입 크기의 변화 없이 입술 모양만 변화시키면서 발성하기 때문에 기존가수 보다 명료성이 낮고 지속시간이 적다고 할 수 있다. 그러므로 랩을 주로 하는 신세대가수보다 기존가수가 말의 의미 전달에 있어서 더 명료하게 전달 할 수 있다.

5. 참 고 문 헌

[1] Dellattre, P.C., A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," Journal of the Acoustical Society of America(JASA) vol. 27, no.4, pp.769-773 1955.

[2] Fant, C. G. M. Acoustic theory of Speech Production. The Hague, The Netherlands : Mouton, 1960.

[3] Flanagen, J. L. Speech Analysis, Synthesis, Perception, 2nd ed. New York : Sprniger-Verlag, 1972.

[4] Lindblom, B. E. F., and J. E. F. Sandberg. "Acoustic consequences of lip, tongue, jaw, and larynx movement," JASA vol.50, pp.1166-1179, 1971.

[5] Peterson, G. E., and H. L. Barney, "Control Methods used in a study of the vowels." JASA vol.24, pp.175-184, 1952.

[6] Stevens, K. N., and A. S. House,. "An acoustical theory of vowel production and some implications," Journal of Speech and Hearing Research, vol.4, 1961.

[7] Zemlin, W., Speech and Hearing Science, Anatomy and Physiology, Englewood Cliffs, N. J., Prentice Hall.

[8] J. L. Flangan, Speech Analysis Synthesis and Perception, 2nd Ed., Springer-Verlag, New York, 1972.

[9] 배명진, 이상효, 디지털 음성분석, 홍영출판사. 1998

[10] 한진수, 음성신호처리, 오성미디어, 2000

[11] 배명진, 디지털 음성합성, 홍영출판사. 1998