한국어 음운체계를 고려한 화자적응 실시간 단모음인식에 관한 연구

황 선 민*, 윤 한 경**, 송 복 희***

Speaker Adapted Real-time Dialogue Speech Recognition Considering Korean Vocal Sound System

Seon-Min Hwang*, Han-Kyung Yun**, Bok-Hee Song***

요 약

음성인식에 관한 연구는 꾸준히 발전되어 다양한 분야에서 제품에 적용되고 있으며, 특히 스마트폰과 차량의 내비게이션 시스템과 같은 정보기기에서의 적용은 매우 활발히 이루어지고 있는 것이 현실임에도 불구하고 음성인식 연구에서 한국어의 음운 특성을 고려한 연구는 미흡한 것도 현실이다. 디지털 콘텐츠에서 일반적으로 립 싱크의 제작은 애니메이터의 수작업을 반드시 필요로 하며, 이는 시간의 투입을 요하는 지루한 작업이다. 립 싱크를 자동 생성하는 응용 소프트웨어를 사용하기도 하나 반드시 수작업으로 수정과 보정 단계가 요구될 뿐만 아니라 영어 기반으로 제작된 립 싱크 자동생성 소프트웨어를 적용하므로 한국어 음운체계가 달라 자연스러운 립 싱크를 얻기 위하여 많은 시간과 노력이 요구된다. 따라서 본 연구에서는 한국어 음운체계를 고려한 가상 오브젝트들의 립 싱크를 자동 생성하기 위한 화자 독립 기반 한국어 단모음 실시간 인식 알고리즘을 개발을 목표로 하며, 그 인식 결과는 립 싱크의 애니메이션 키로 활용하고자 한다.

ABSTRACT

Voice Recognition technique has been developed and it has been actively applied to various information devices such as smart phones and car navigation system. But the basic research technique related the speech recognition is based on research results in English. Since the lip sync producing generally requires tedious hand work of animators and it serious affects the animation producing cost and development period to get a high quality lip animation. In this research, a real time processed automatic lip sync algorithm for virtual characters in digital contents is studied by considering Korean vocal sound system. This suggested algorithm contributes to produce a natural lip animation with the lower producing cost and the shorter development period.

keywords speaker adaptive dialogue recognition, FFT, single vowel recognition, lip sync

* 한국기술교육대학교 컴퓨터공학부,(snhwang@kut.ac.kr),
** 교신저자 : 한국기술교육대학교 컴퓨터공학부 교수 (hkyun@kut.ac.kr),
*** 한국기술교육대학교 디자인공학과 교수
접수일자 : 2013년 10 월 07 일, 수정일자 : 2013년 11 월 08 일, 심사완료일자 : 2013년 12 월 09 일

Ⅰ. 서 론

음성인식에 관한 연구는 꾸준히 발전되어 다양한 분야에서 제품에 적용되고 있으며, 특히 스마트폰과 차량의 내비게이션 시스템과 같은 정보기기에서 적극적인 적용이 매우 활발히 이루어지고 있다. 그럼에도 불구하고 음성인식 연구에서 적용되는 기본적인 방법은 영어에서 활용하던 방법들을 적용함으로써 기존의 많은 음성인식 연구는 한국어의 음운 체계와 특성이 간과된 경우가 대부분이다. 또한, 립 싱크의 자동 생성을 위한 국내 연구에서도 일반 적인 음성 인식 방법론을 적용하여 접근함으로써 그 결과가 범용적으로 활용된 사례는 없는 실정이다. 따라서 가상 오브젝트의 립 싱크 제작과정과 실정을 고려하고 한국어 음운 체계와 특성을 고려한 립 싱크 애니메이션을 자동 생성을 위한 연구가 필요하다.

디지털 콘텐츠에서 가상 오브젝트의 립 싱크 애니메이션과 대사의 불일치는 사용자들의 현실감을 떨어뜨려 몰입을 방해하는 요소로 작용하여 재미를 감소시키고 이러닝에서 학습효과를 저해시키는 요인이 된다. 따라서 사용자들이 느끼는 부자연스러운 립 애니메이션을 자연스럽게 수정 보완하기 위하여 애니메이터들의 수작업을 거치게 되며 이는 개발 기간과 비용의 증가 요인이 될 뿐만 아니라 매우 지루하고 반복적인 작업으로 참여자들의 생산성은 낮아지므로 고품질의 립 애니메이션은 개발 비용의 중요한 증가 요인 중에 하나이다. 그러나 현실적으로 제작 스튜디오 측에서 개발 비용의 추가를 요청하기는 매우 어려운 실정이다. 따라서 콘텐츠의 사용자들이 용인하는 정도의 립 싱크 애니메이션의 제작이 가능한 솔루션을 개발 되여야만 한다. 이를 위하여 한국어 음운 체계에 따른 발성 특성을 조사 분석이 이루어져야 할 뿐만 아니라 한국어 대화 특성을 조사 분석이 이루어 져야 한다. 또한, 립 싱크 애니메이션을 사용자들이 용인하는 정도를 정량화하기 위하여 애니메이션 속성에 따른 립 싱크 애니메이션을 분석할 필요가 있다.

디지털 콘텐츠에서 대화나 설명은 특정화지인 성우들이 참여하므로 개인차와 무관한 화자 종속 음성인식이라고 판단할 수 있으며 개인의 생리적 조건과 환경의 영향을 최소화하기 위하여 녹음을 시작하기 전에 참조 패턴을 발성하여 특징량을 추출하기로 한다. 한글의 음운체계에서 모든 음소들은 각각 고유의 음가를 갖고 있으며 일정한 발음 규칙을 적용하여 발성되므로 음성 인식 측면에서는 상대적으로 영어보다 용이하다고 판단할 수 있다.

음성인식연구의 결과를 보면 대부분의 에너지가 모음에 분포되어 있으므로 모음의 입 모양이 립 싱크에서 중요한 요소임을 알 수 있으며 실질적으로 사람의 입모습도 동일하다. 따라서 립 싱크 애니메이션의 중요한 요소는 모음 인식이다. 모음의 발성은 입술모양과 혀의 위치와 높이에 따라 결정이 되나 애니메이션에서 통상적으로 입술의 여는 형태와 시간으로 나타나지만, 고품질이나 3D 립 싱크 애니메이션에서는 혀의 높이도 필요하다.

제안된 알고리즘은 앞에서 기술된 모든 특성과 제약조건을 고려하여 모음 중에서 단모음을 중심으로 음성인식이 실시간으로 인식될 수 있도록 경량화된 솔루션으로 실시간 처리가 가능할 뿐만 아니라 최적 시스템으로 간단히 구성함으로써 자본과 형편에 여유가 없는 기존의 애니메이션 제작 스튜디오들도 큰 부담이 없도록 개발되었다.

Ⅱ. 한국어 발성 특성

1. 모음의 음운 체계

영어와 달리 한글 모음의 발음은 항상 일정하며, 모음들은 그 구성 체계에 따라 단모음과 복모음 또는 이중모음으로 크게 구분된다. 단모음은 입술의 모양에 따라 발음할 때 둥글게 오므리지 않는 평순 모음과 입술을 둥글게 오므려 발음하는 원순모음으로 구분하며, 혀의 위치에 따라 혀의 앞쪽에서 발음되는 전설모음과 혀의 뒤쪽과 입천장 뒤쪽의 부드러운 부분에서 소리 나는 후설모음, 입을 조금 열고 혀의 위치를 높여 발음하는 고모음(폐모음), 입을 보통으로 열고 혀의 높이를 중간으로 하여 발음하는 중모음, 입을 크게 벌리고 혀의 위치를 가장 낮추어서 발음하는 저모음(개모음)으

로 구분된다. 한국어의 모음 별 입술 모양과 혀의 위치 및 높이는 표 1과 그림 1와 같으며, 복모음은 단모음의 조합으로 표시되며 소리를 내는 도중에 입술 모양이나 혀의 위치가 시간에 따라 달라지는 모음이다.

표 1. 한국어 단모음별 입술모양과 혀의 위치 및 높이
Table 1. Lip Shape and Tongue's Podition of Korean Single Vowel

혀의 최고점 위치		앞(전설모음)		뒤(후설모음)	
입술모양		평평한 입술 모양 (평순 모음)	둥근 입술 모양 (원순 모음)	평평한 입술 모양 (평순 모음)	둥근 입술 모양 (원순 모음)
혀의 높이	H(고모음)	ㅣ	ㅟ	ㅡ	ㅜ
	M(중모음)	ㅖ	ㅚ	ㅓ	ㅗ
	L(저모음)	ㅐ		ㅏ	

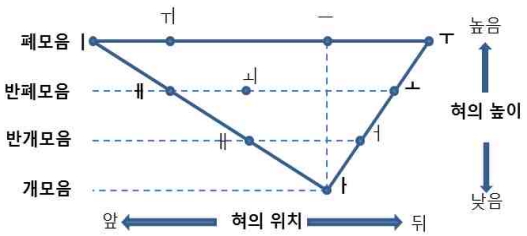

그림 1. 모음 삼각도
Fig. 1 Korean Single Vowel Triangle

디지털 콘텐츠 에서 가상 캐릭터 또는 오브젝트의 립싱크는 우선적으로 입의 모양을 음절단위에 동기화하는 것이 중요하며 정밀한 입술 모양도 자연스러운 애니메이션을 위하여 필수적이다. 입의 모양은 입술모양에 따라 평평한 모양과 둥근 모양, 혀는 위치와 높이에 따라 결정된다. 따라서 디지털 콘텐츠를 제작함에 있어 시각적 효과 측면과 제작상의 효율적 측면을 고려한다면 전신캐릭터의 립싱크는 입술만으로 싱크를 맞추어도 자연스러운 시각적 효과를 얻을 수 있으나 캐릭터의 두상이 클로즈업 된 경우에는 입술모양뿐만 아니라 혀의 위치 및 높이가 고려되어야 한다.

2. 한국어 대화의 특성

한국어의 통상적인 읽기 평균 속도는 348 SPM(Syllables Per Minute)이며 빠른 읽기속도는 426 SPM 정도로 알려져 있으며 말하기 속도는 평균 265SPM이고 최저 118SPM에서 409SPM이라고 보고되었다[1]. 즉, 평균 대화속도는 265SPM이므로 초당 약 4.4 음절이 발성되어 음절 당 약 227msec 가 된다. 이는 초당 30프레임 애니메이션에서 6.8프레임에 해당한다. 한국어 대화에서 최대 속도는 400 SPM 정도이므로 초당 6.7~6.8 음절들이 발성되어 8KHz의 샘플주파수를 사용하면 음절 당 샘플수는약 1200여개가 된다. 또한, 최저 대화속도는 118SPM이므로 초당 약 2음절이 발성되어 8KHz의 샘플주파수를 사용하면 음절 당 샘플수는 약 4000여개가 된다.

Ⅲ. 실험

본 연구에서 제안하는 모음 인식 알고리즘은 성우의 발성에 따른 캐릭터의 입모양의 애니메이션을 자동화하는 것이 목적이므로 발성에 따른 자연스러운 입모양을 자동적으로 구현하여 수작업에서 발생하는 미스 매치의 오류를 최소화하고 자연스러운 입모양을 생성하기 위한 것이다.

1. 실험 방법

지금까지 음성 인식을 이용해 컴퓨터와 인간 사이의 의사소통을 위한 연구가 활발히 진행되어 왔다. 대표적인 음성 인식 기법으로 멜 주파수 켑스트럼 분석 (Mel Frequency Cepstrum Analysis)[2]과 선형 예측 (Linear Predictive)[2] 기법들이 일반적으로 사용되어 왔다. MFC 해석은 잡음 환경에 강한 특성을 가지나 속도가 LP 해석에 비하여 느린 단점이 있으며, LP 해석은 속도가 MFC 해석에 비해 빠르나 잡음 환경 하에서 인식률이 떨어지는 단점을 보인다.

음성 인식을 위한 기본적인 절차는 음성 데이터 입력, 시간-주파수 변환, 특징 벡터 추출, 인식으로 구성된다. 대부분의 연구에서 특징벡터를 추출하는 부분이 인식률에 많은 영향을 끼치는 것으로 알려져 왔다. 음성 데이터를 FFT에 의해 주파수 도메

인으로 변환하면 특정 구역에서 최대값을 가지는 형태의 주파수 특성을 보이는데 이러한 특성을 포만트(formant)라 한다[3]. 포만트는 위치에 따라 F1, F2, F3, F4 로 구성된다.

한국어의 모음은 단모음과 단모음의 병합으로 이루어진 복모음이므로 입의 모양과 혀의 위치 및 높이는 단모음들의 기본 형상에 종속되어, 입의 모양은 평순모음과 원순모음, 혀의 높이에 따른 개모음, 반개모음 또는 반폐모음, 폐모음으로 구분되고 혀의 위치에 따른 전설모음과 후설 모음으로 분류된다. 사전 연구에 의하면 입의 형상은 열린 정도에 따라 열린, 중간 열린, 닫힌 입술과 혀의 높이에 따라 높은, 중간, 낮은 혀들의 조합으로 디지털 콘텐츠에서 자연스러운 립 싱크 애니메이션이 가능하였다. 따라서 실시간 처리를 위하여 시스템의 단순화가 관건이므로 캐릭터의 입모양을 '아', '에', '이', '오', '우' 의 단모음 5개로 한정하여 성우의 발성에 따른 모음을 인식하여 3개의 입모양과 3개의 혀의 위치의 조합으로 제한된 립싱크 애니메이션 제작이 가능한 인식 알고리즘을 제안한다.

본 알고리즘은 첫째, 마이크를 통해 '아', '에', '이', '오', '우' 의 단모음 다섯 개를 3번 반복 입력받아 FFT 과정을 거쳐 모음별 기준 F1과 F2 포만트들을 추출한다. 둘째, 실시간 음성 데이터를 입력받아 FFT 과정을 거쳐 모음 포만트를 추출한다. 셋째, 기준 포만트들과 실시간 추출된 모음 포만트의 F1, F2 주파수를 비교하여 모음을 인식하며, 그 결과는 립 싱크를 위한 립 애니메이션의 키로 사용된다. 전체 과정[4]은 그림 2와 같으며 본 연구에서는 실시간 모음 인식으로 제한한다.

본 연구에서 음성 입력에 사용된 마이크는 소규모 제작사들이나 개인에 부담을 최소화하기 위하여 범용적인 것으로 SAMSON사의 Go Mic Portable USB 제품을 사용하였다. 입력된 음성 신호는 PCM 과정을 통해 디지털 데이터로 변환된다. PCM 과정은 샘플링, 양자화, 부호화 과정으로 나눠진다. 본 연구에서는 디지털 데이터 추출을 위해 1채널, 16비트, 8KHz의 주파수로 샘플링하였다. PCM 과정에 의해 변환된 음성 디지털 데이터의 샘플은 그림 3과 같다.

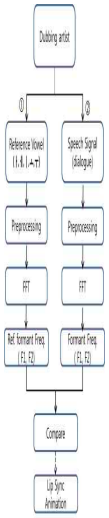

그림 2. 한국어 단모음 인식 알고리즘
Fig. 2. Procedure for the real-time Korean single vowels recognition algorithm

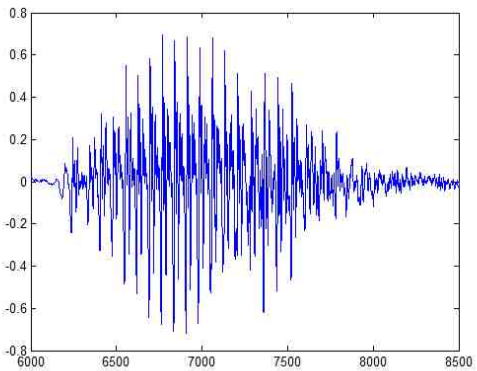

그림 3. 음성 입력 신호의 표본 데이터
Fig. 3. Sampled data of voice input

모음 별 포만트 주파수 F1과 F2의 추출은 그림 3과 같은 디지털 PCM 데이터를 1024 FFT로 처리한 후 전 구간을 40Hz구간으로 나누어 각 구간의 진폭의 합을 구하여 초기의 최고값을 F1, 두 번째 최고값이 F2가 된다. 40Hz의 구간은 실시간 처리를 위하여 실험값으로 결정하였으며 기존에 보고된 포만트 값을 참조하여 1Khz 이상에서 최대값이 생성되면 모음이 아닌 것으로 간주하였다.

인식률을 향상시키기 위하여 5개의 기본 단모음들의 F1과 F2를 성우들이 대사를 녹음하기 직

전에 추출하여 참조 데이터를 매번 생성하여 개인의 생리적 조건이나 환경에 의한 변화를 최소화하였으며, 전형적인 결과를 그림 4에 보였으며 이는 평평한 입술모양(평순모음)과 중간 정도의 혀 높이(중모음)인 단모음 'ㅔ'의 FFT 해석 결과이며 좌측에 표시된 세로 막대 구간이 F1이고 다음 막대 구간이 F2이며 'ㅔ'의 참조 데이터가 된다.

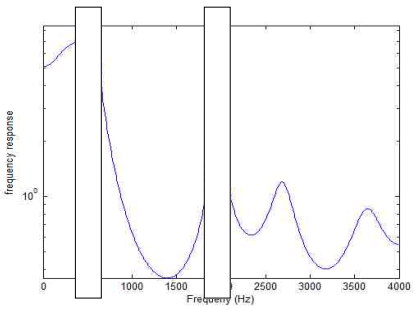

그림 4. 단모음 '에' 의 기준 포만트 주파수
Fig. 4. Reference formant frequencies of the single vowel 'ㅔ/[e]'

Ⅳ. 실험 결과

본 연구에서 제안한 방법을 사용하여 성인 남성의 음성입력에서 실시간 모음 인식 실행하여 'ㅔ'로 인식된 실험 데이터를 분석하면 그림 5와 같으며 'ㅔ'의 참조 데이터와 실시간으로 추출된 음성 입력신호의 F1과 F2 대역이 동일함을 알 수 있다.

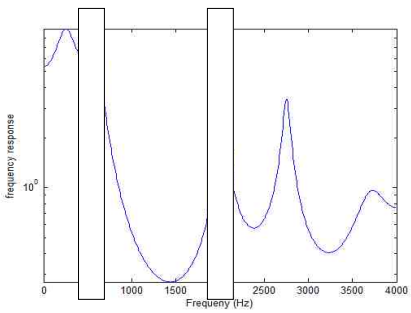

그림 5. 단모음 '에' 로 인식된 실시간 음성 입력신호
Fig. 5. Real time voice input recognized as the single vowel ' ㅔ/[e]'

본 연구에서 제안한 방법을 사용하여 특정한 남성 성우 3인의 모음 인식 결과는 표 2와 같다. 모음 '에'의 인식률이 가장 좋았으며 '오' 와 '우'의 인식률이 다른 모음의 인식률에 비해 낮은 이유는 입술 모양이나 혀의 위치는 동일하고 단지 혀의 높이만 차이가 나므로 'ㅗ'와 'ㅜ'는 상호 오인식에 기인한 것으로 분석되었다.

표 1. 제안된 알고리즘의 한국어 단모음 인식률
Table 1. Korean single vowel recognition rate of the proposed algorithm

한국어의 단모음	인식률(%)
'ㅏ' [a]	98.5
'ㅔ'[e]	98.6
'ㅣ'[i]	97.2
'ㅗ'[o]	96.3
'ㅜ'[u]	96.1
평균.	**97.3**

한국어의 음절문자로 초성 중성 종성 들의 조합으로 음절이 이루어지고 초성 및 종성들이 발성 중에 중성 인 모음에게 주는 영향을 분석한 결과를 그림 6에 보였으며 이는 음성신호를 PCM 데이터를 전처리 없이 FFT 주파수 분석을 한 결과이다. 평순,후설,저 모음인 'ㅏ'는 자음의 파열음, 마찰음 및 파찰음의 '예사소리'와 파열음의 거센소리 및 연구개음 중 비음'ㅇ'에서 F1과 F2 포만트 대역이 형성되었으며, 파열음의 된소리와 마찰음 예사소리 중 성문음과 경구개음 중 파찰음의 거센소리 및 비음의 양순음 치조음에서 F1과 F2 주파수 대역이 확장되었다.

평순,전설,중 모음인 'ㅔ'는 연구개음, 성문음, 비음 및 유음에서 F1과 F2 포만트 대역이 형성되었으며 양순음과 치조음 중 파열음의 예사소리와 거센소리는 F1 포만트 대역이 확장되고 F2는 대역이 분리되는 현상을 보였다.

평순, 전설, 고모음인 'ㅣ'는 '연구개음, 성문음, 비음 및 유음과 양순음과 연구개음 중 파열음의 거센소리에서 F1 포만트 대역이 형성되고, 파열음에서 양순음과 치조음에서 대역의 확장 현상을 보였으며 F2 포만트 대역은 파열음 예사소리 중 치조음과 연구개음, 파열음의 거센소리 중 양순음과 비음 중 치조음에서 형성된다.

원순, 후설, 중모음인 'ㅗ'의 F1 주파수 대역은

파열음의 예사소리 중 양순음과 연구개음과 마찰음의 예사소리 중 치조음, 파찰음의 거센소리 중 경구개음, 비음과 유음중 치조음에서 F1 포만트 대역이 형성되었으며, F2 포만트 대역은 파열음의 예사소리 중 치조음과 연구개음과 파찰음의 예사소리인 경구개음에서 형성되었고 파열음의 거센소리 중 양순음과 연구개음, 마찰음의 예사소리 중 성문음파찰음의 거센소리인 경구개음들에서 F2 포만트 대역이 분리되는 현상을 보였다.

원순, 후설, 고모음인 'ㅜ'의 F1 주파수 대역은 파열음의 예사소리인 양순음, 치조음, 연구개음과 파찰음의 예사소리인 경구개음, 유음과 비음에서 F1 포만트 대역이 형성되었고 파열음의 거센소리와 마찰음의 예사소리, 파찰음의 거센소리에서 대역의 확장 현상을 보였으며 F1과 동일하게 주파수 대역이 형성되었으며 파열음의 거센소리 중 양순음과 연구개음, 파찰음의 거센소리인 경구개음들에서 F2 포만트 대역이 분리되는 현상을 보였다.

자음 음소들에 따른 포만트 주파수의 영향을 분석하여 관계를 규명한다면 범용적인 참조 데이터 패턴을 간단하 얻을 수 있어 제시된 알고리즘을 보다 경량화면서 인식률 향상이 가능할 것이다.

Ⅴ. 결 론

본 연구에서는 대사에 동기화된 캐릭터의 입모양을 자동 생성하기 위한 것으로 시각적으로 자연스러운 입모양의 구현을 위하여 음성인식을 적용한 것이다. 따라서 한국어의 대표 단모음이며 모음 삼각도에서 입을 벌린 정도와 혀의 위치에 따라 아(평순 저모음), 에(평순 중모음), 이(평순 고모음), 오(원순 중모음), 우(원순 고모음)로 구분하여 자연스러운 립 싱크가 가능하도록 실시간 처리가 가능한 한국어 단모음 인식 알고리즘을 개발하였다. 이는 모음 인식 시스템의 실시간 처리를 위하여 애니메이션의 속성과 한국어의 발성체계를 활용하여 F1과 F2 포만트 주파수만을 특징량으로 사용하여 알고리즘을 경량화하였다. 또한 음절에서 단모음의 포만트 추출이 가능하다면 참조 데이터의 패턴을 제시함으로써 인식 알고리즘은 더욱 간단해짐으로써 처리속도의 향상이 기대된다.

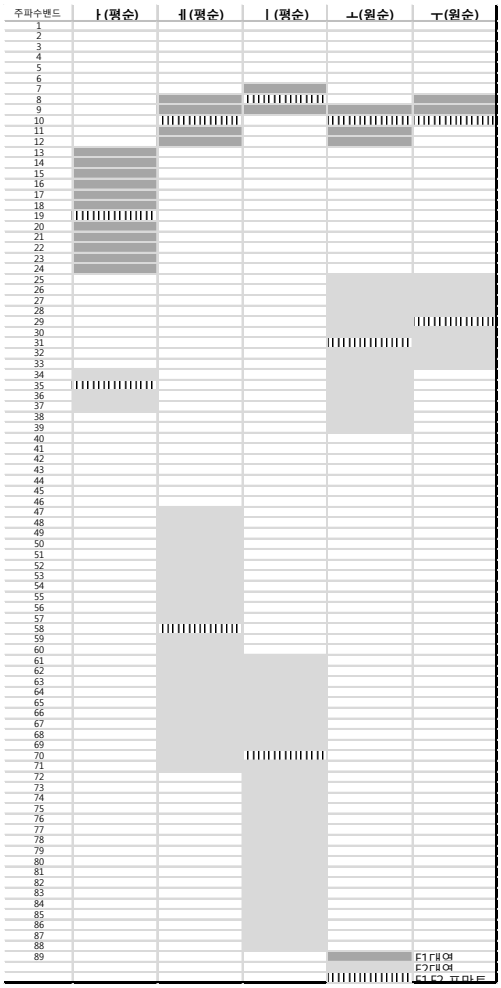

그림 6. 자음의 영향을 고려한 음절의 포만트 주파수 분석

Fig. 6. Formant analysis of syllables included a single vowel

참 고 문 헌

[1] Moonja Shin, and Sook-ja Han, "A Study of Rate and Fluency in Normal Speaker", Speech Science, vol.10, no.2, pp.159-168, 2006.

[2] J. D. Markel, and A. H. Gray. Linear Prediction of Speech, Springer Verlag, 1976.

[2] J. D. Markel, and A. H. Gray. Linear

Prediction of Speech, Springer Verlag, 1976.

[3] D. H. Bailey, and P. N. Swarztrauber. A Fast Method for Numerical Evaluation of Continuous Fourier and Laplace Transform, Journal on Scientific Computing, vol. 15, no. 5, pp. 1105-1110, Sep. 1994.

[4] S. M. Hwang, H. K. Yun, and B. H. Song. Automatic Lip Sync Solution for Virtual Characters in 3D Animations, ICCT 2013, pp.432-433, Chiang Mai 2013.

저자약력

황 선 민(Sun-Min Hwang) 정회원

2013 현재
한국기술교육대학교
정보미디어연구소
선임연구원

<관심분야 > 신호처리, 전자통신

윤 한 경(Han-Kyung Yun) 종신회원

2013 현재
한국기술교육대학교
컴퓨터공학부 교수

<관심분야 > 인공지능, HCI Haptic

송 복 희(Bok-Hee Song) 정회원

2013 현재
한국기술교육대학교
디자인공학과 교수

<관심분야 > 산업디자인, 디자인경영, UI, UX