대역별로 여과한 음성 강도의 차이값과 상관계수에 의한 화자확인 연구*

A Study on Speaker Identification by Difference Sum and Correlation Coefficient of Intensity Levels from Band-pass Filtered Sounds

양 병 곤**
Byunggon Yang

ABSTRACT

This study attempted to examine a speaker identification method using difference sum and correlation coefficient determined from a pair of intensity level matrices of band-pass-filtered numeric sounds produced by ten female speakers of similar age and height. Subjects recorded three digit numbers at a quiet room at a sampling rate of 22 kHz on a personal computer. Collected data were band-pass-filtered at five different band ranges. Then, matrices of five intensity levels at 100 proportional time points were obtained. Pearson correlation coefficients and the sum of absolute intensity differences between a pair of given matrices were determined within and across the speakers. Results showed that very high correlation coefficient and small difference sum generally occurred within each speaker but some individual variation was also observed. Thus, the matrix pair with a higher coefficient and a smaller difference sum was averaged to form each individual's model. Comparison among the speakers yielded generally low coefficients and large differences, which suggests successful speaker identification, but among them there were a few cases with very high coefficients and small differences. Future studies will focus on finer band ranges and additional spectral parameters at some peak points of the intensity contour at a low frequency band.

Keywords: Speaker Identification, Individual Variation, Band-pass Filtering

1. 머릿말

인터넷을 통한 화자확인 과정은 먼저 등록할 사람이 여러 번 발성한 목소리의 음향적 특징을 추출한 뒤, 이를 모델로 설정해두고, 나중에 동일한 사람이 접속하여 등록된 단어를 발음했을 때, 음향적으로 분석한 결과가 이 사람의 목소리와 일치하는지 여부를 확인하는 방법이다. 사람의 목소리에는 태어난 지방, 성질이나 사회의 계층 등의 '지시적인 정

* 본 연구는 한국과학재단 목적기초연구(R01-1999-000-00229-0)지원으로 수행되었음.
** 동의대학교 영어영문학과

보' (Abercrombie, 1967), 또는 '개인적인 정보' (Ladefoged 외, 1957)가 포함되어 있다. 이런 정보를 통해 우리는 쉽게 화자의 정체와 신체적이고 정서적인 특징을 읽을 수 있고 목소리만으로도 화자를 알아 맞출 수 있다 (Van Lancker 외 1985a;1985b). 지금까지의 화자확인 연구에서는 화자마다 발음한 음성에 개인만의 독특한 음향적 특징이 있을 것이라 미리 가정하고 이를 정확히 측정하는 방식에 대한 연구를 추진해 왔다. 하지만, 기존 연구에서도 밝혀졌듯이, 측정 음성분석 소프트웨어의 파라미터 설정에 따라 많은 차이가 나타나게 되며, 소음에 의한 측정값의 에러를 극복하기 어렵거나, 여성화자일 때는 매우 불안정한 값들이 구해져서 이를 모델로 설정하는데 어려움이 많았다 (양병곤 2000, 2001;양병곤 외 2002). 결국, 화자의 말투에는 개인마다의 특색이 존재하므로 이를 너무 세밀하게 분석하게되면 개인별로도 여러 가지로 변하는 값들 때문에 모델을 세우기 어렵고, 포먼트, 피치 등과 같은 음성언어학적인 부분의 정보만으로는 다른 화자와 동일시될 확률이 높다. 특히, 기존의 음성분석소프트웨어는 여성화자의 발성에 대해 측정상의 오류가 많기 때문에 이를 근거로 모델을 설정하기가 어려움이 많다. 따라서, 화자확인과정의 연구에서, 측정 에러가 적으면서도 적절한 해상도의 분석파라미터를 이용하는 것이 아주 중요한 과제로 여겨진다. 특히, 시간적으로 단절된 부분의 주파수축선상에 분석에 치중해왔던 접근방법을 시간축의 역동적인 정보도 반영할 수 있는 방법의 모색이 필요하다.

이 연구의 목적은 앞서 지적한 연구방향을 검토하기 위해 기존의 파라미터 분석방식보다는 에러가 적은 대역별로 여과한 뒤 나온 강도값을 전체지속시간에 비하여 각 지점마다 구하여 화자 개인별 모델을 설정하고, 이 모델간의 상호 비교를 통해 화자확인을 할 수 있는 방법을 탐구해 보기로 한다. 이러한 목적을 달성하기 위해, 거의 비슷한 신체조건을 가진 10명의 여성화자가 임의의 순서로 발성한 세자리 숫자음을 5개의 대역별로 여과시킨 뒤 각 대역에 대한 강도값의 행렬을 구하고, 각 화자별로 다섯 번씩 발음한 숫자음의 상관계수와 절대차이합을 구하여 화자내 변이를 통계적으로 살펴보고, 이들 비교 쌍 가운데 가장 상관도가 높은 두 개의 발음의 평균행렬을 구하여 각 화자의 모델로 설정한다. 이어서, 이들 모델행렬을 각 화자끼리 비교하여 상관계수와 절대차이합을 구하여 서로 비교해 보고자 한다. 이러한 방법은 기존의 분석프로그램의 한계점을 극복하면서도 동시에 어느 정도 세밀한 정보를 구하여 화자간에 비교함으로써 화자확인에 필요한 소프트웨어 개발에 도움이 될 것이다.

2. 연구 방법

2.1 피험자 녹음과정

피험자는 동의대학교에 재학하는 건강하고 청각에 이상이 없는 여학생 10명을 임의로 선정했다. 표 1은 화자의 나이와 키를 나타내주고 있다. 피험자의 나이는 평균 21세이었고, 키의 평균은 162.2 cm이었다. 거의 비슷한 나이와 키를 가진 화자를 선택한 이유는 측정값의 에러가 많고, 비슷한 음향적 특성을 지니는 여성화자의 처리과정이 잘 된다면 남성이나 신체적 조건이 다른 피험자들의 구별은 보다 쉬울 것으로 여겨졌기 때문이다. 피

험자들에게는 인터넷을 통해 자신의 목소리를 저장하고 확인하는데 사용한다는 실험의 목적을 간단히 설명하여 줌으로써 화자내에서는 대체로 동일한 목소리로 발음하도록 유도했다. 이렇게 화자내의 변화율이 적을수록 다른 화자와 구별할 때 보다 좁은 통과범위를 확보할 수 있을 것이다. 음성자료는 조용한 연구실에서, G4노트북 컴퓨터에 Sound Studio라는 음성녹음 프로그램을 제어하는 스크립트를 작성하여 수집했다. 각 피험자는 컴퓨터 화면에 자동으로 임의의 순서로 된 숫자가 나타나는 것을 보고 4초 이내에 발음하면, 컴퓨터가 그 음성을 재생시킨 뒤, 피험자가 만족하면 저장단추를 누르고 잘못된 발음이면 다시 녹음할 수 있도록 하였다. 음성표본속도는 22 kHz, 단음으로 저장했다.

표 1. 피험자 정보

화자	나이(세)	키(cm)	화자	나이(세)	키(cm)
s1	21	164	s6	21	163
s2	21	165	s7	21	160
s3	21	158	s8	21	168
s4	21	164	s9	21	162
s5	21	160	s10	19	158

2.2 자료분석

피험자들은 다양한 숫자음 조합을 녹음하였으나, 이 논문에서는 "678", "789"과 "891"로 한정하여 분석했다. 이런 숫자음을 선택한 이유는 모음삼각도에서 모서리 음에 해당하므로 개인별 성도의 특성이 보다 많이 반영될 것이고 "육백 칠십 팔" 등의 발음으로 시간축에서 역동적인 변화를 포착할 수 있을 것으로 여겨졌기 때문이다. 분석과정은 실제 소프트웨어로 구현할 수 있도록 프라트 스크립트를 작성하여 자동화시켰다. 먼저, 4초 이내에 발성한 음성파형의 전체 강도값의 평균과 표준편차를 구하여 (평균+1/5표준편차)에 해당하는 강도의 문턱값을 음성의 처음과 끝에서 각각 찾아내어 시작점과 끝점을 지정했다. 이어서, 시작점과 끝점 사이에 해당하는 음성 부분을 따로 뽑아내어 200 Hz에서 4200 Hz 사이를 800 Hz간격으로 다섯 개의 대역으로 분리하였다. 실제 다섯 개 대역보다 더 세분하지 않은 것은, 저자가 다섯 개로 분리한 대역을 동시에 재생하며 들어봤을 때 원음과 별로 차이가 없음를 확인하였고 또한 대역이 많을수록 개인별 모델의 안정성에 어려움이 있을 것으로 예상했기 때문이다. 대역별로 구분된 음성파형의 강도값을 200 Hz간격으로 추출하여 전체지속시간을 100등분하여 각 지점의 강도값을 구했다. 이때 강도가 일정수준 이하 (평균-1/5표준편차)이면 동일한 값을 부여하였다. 수집된 음향적 변수로는 모두 약 75,000개의 자료 (10명X5번씩 발음X3개의 숫자음X100개의 시간점X5개의 대역별 강도자료)를 구했다. 음성 비교는 화자내 비교와 화자간 비교로 나누어진다. 화자내 비교는 각 개인마다 다섯 번의 발음을 서로 비교하여 상관계수와 절대차이합의 차이를 구하였다 (양병곤 2002). 이어서, 비교된 쌍의 상관계수가 아주 높으면서도 절대차이합이 적은 쌍을 개인별로 선정하여 이 두 행렬의 각 항의 평균을 구하여 화자 행렬모델로 지정했다. 화자간 비교에서는 각 화자별 모델을 서로 비교하여 상관계수와 절대차이합을 구하였다.

3. 분석 결과와 토론

3.1 화자내 비교

먼저 각 화자별로 발성한 다섯 번의 숫자음에 대한 상호 비교 결과를 살펴보자. 표 1은 s1(1번 피험자)의 다섯 번의 발음의 상호 비교 결과를 나타내어주고 있다.

표 1. s1이 다섯 번 발음한 숫자음 "678"의 상호 비교 상관계수와 절대차이합. t1은 첫 번째 발음을 나타낸다.

	t2	t3	t4	t5
t1	0.89	0.85	0.95	0.88
t2		0.92	0.90	0.95
t3			0.88	0.97
t4				0.90
t1	1962	2423	1532	1841
t2		1646	1543	1496
t3			1794	1805
t4				2101

표 1을 보면 s1의 발음은 상관계수로는 세 번째와 다섯 번째의 발음이 가장 높게 나타난 0.97이고, 절대차이합은 두 번째와 다섯 번째의 비교치가 가장 적은 1496을 기록하고 있다. 화자 개인별 숫자음 발음의 변화 정도를 일목요연하게 파악하기 위해, 각 숫자음별로 모아 상관계수와 절대차이합의 평균과 표준편차를 그래프로 그려보면 그림 1에서 그림 3과 같다.

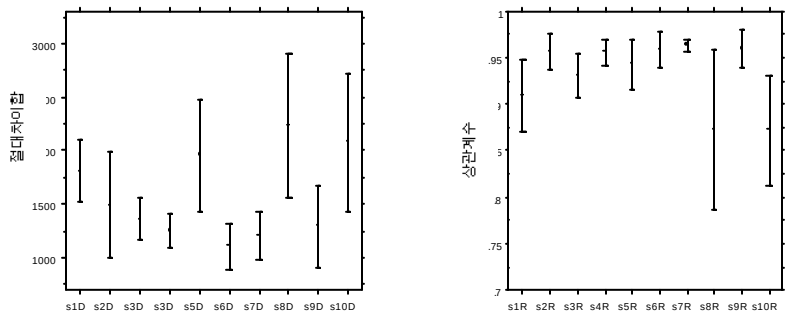

그림 1. 숫자음 "678"에 대한 화자별 발음의 상관계수와 절대차이합의 산포도. 그림에서 각 선의 중심점은 개인별 평균값을 나타내며 아래 위 1표준편차를 나타내었다. s1R은 제1피험자의 상관계수를 나타내고 s1D는 절대차이합을 의미한다.

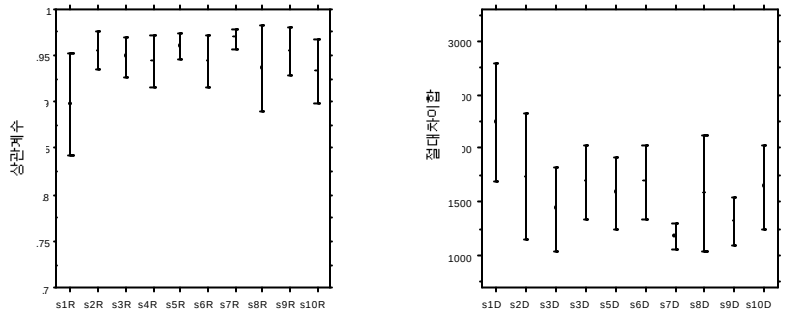

그림 2. 숫자음 "789"에 대한 화자별 발음의 상관계수와 절대차이합의 산포도

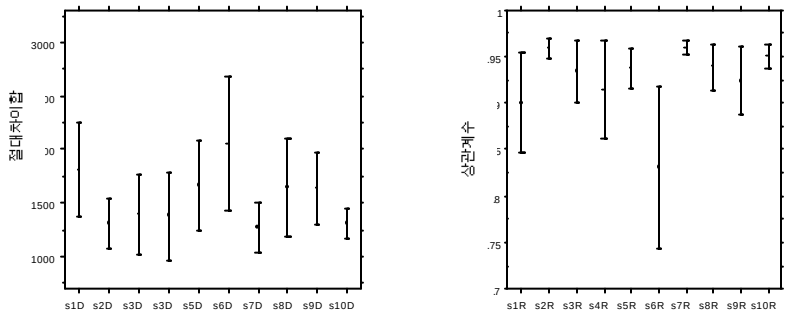

그림 3 숫자음 "891"에 대한 화자별 발음의 상관계수와 절대차이합의 산포도

이 그림들의 통계적인 특성을 살펴보면 숫자음 "678"에서는 상관계수의 범위가 0.72~0.98이고 절대차이합은 738~3242에 분포한다. 평균 상관계수는 0.93이며 절대차이합의 평균은 1581이었다. 숫자음 "789"에서는 상관계수의 범위가 0.79~0.98이고 절대차이합은 909~3091에 분포한다. 평균 상관계수는 0.94이며 절대차이합의 평균은 1610이었다. 마지막으로 숫자음 "891"에서는 상관계수의 범위가 0.71~0.98이고 절대차이합은 804~2867에 분포한다. 평균 상관계수는 0.93이며 절대차이합의 평균은 1550이었다.

이 그림의 자료 분포에서 살펴보면, 숫자음별로 개인별로 매우 안정적인 화자가 있고 불안정한 화자도 있음을 알 수 있다. 해당 숫자음내에서는 대체로 화자간의 변화범위가 대체로 비슷하지만, 다른 두 숫자음보다는 숫자음 "789"에서 상관계수가 대체로 0.9이상의 분포가 많다. 개인별로는 s2와 s7이 모든 숫자음에서 매우 높은 상관계수를 보이고 있으며, s6과 s8은 숫자음마다 계수가 달라 안정성이 떨어진다.

절대차이합은 작을수록 개인별 발음이 안정적임을 보여주는데, s7이 대체로 모든 숫자음에서 작게 나타나 있고, s1과 s8은 변화율이 적은 숫자와 높은 숫자로 나뉘어진다. 따라서, 화자개인별로 항상 동일한 음향적 특징을 가지는 음성을 산출하기 어렵기 때문에 각 화자의 발음에서도 일정한 기준이하의 안정적이지 못한 발음들은 화자확인의 모델을 만들

때는 제외시키는 것이 바람직할 것으로 여겨진다. 실제 모델이 설정된 이후에 접속하는 화자확인과정에서 들어오는 데이터는 지정한 문턱값 이상의 기준에 맞을 때만, 이미 구축해놓은 모델을 수정하는데 사용한다면, 모델 설정이후에도 지속적으로 변해 가는 더 정밀한 화자확인 기준을 마련하는데 도움이 될 것이다.

3.2 화자간 비교

앞 절에서 화자마다 다섯 번 발음한 숫자음 가운데 개인별 변화가 있음을 확인했다. 따라서, 이 연구에서는 화자의 발음이라면 모두 동일한 목표음을 발음한 것으로 간주하고 평균모델을 설정하는 기존의 방법과는 달리 다섯 번의 발음에 대한 상호 비교에서 보다 높은 상관계수와 적은 절대차이합을 보이는 쌍의 평균을 개별모델로 설정하여 화자간의 비교를 시도해 보았다. 표 2에서 표 7까지는 각 숫자음별로 화자간 비교를 한 뒤 상관계수와 절대차이합을 각각 나타내어준다.

표 2. 숫자음 "678"의 각 화자별 모델간의 상관계수. s1은 제1피험자를 나타낸다.

	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	0.69	0.88	0.71	0.84	0.44	0.92	0.48	0.89	0.82
s2		0.60	0.86	0.83	0.77	0.79	0.79	0.67	0.75
s3			0.61	0.77	0.38	0.86	0.47	0.91	0.76
s4				0.86	0.7	0.77	0.65	0.67	0.72
s5					0.63	0.88	0.65	0.78	0.81
s6						0.53	0.89	0.42	0.63
s7							0.60	0.87	0.83
s8								0.51	0.62
s9									0.80

표 3. 숫자음 "678"의 각 화자별 모델간의 비교에 따른 절대차이합.

	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	3640	3409	3430	2942	4777	3499	4436	4266	6111
s2		4089	2471	2590	3132	3337	3046	4410	5353
s3			3529	2900	4656	2032	4928	1877	3881
s4				2045	3114	2937	3846	3874	4788
s5					3710	2324	3990	3443	4313
s6						4397	2281	5163	5729
s7							4732	2105	3420
s8								5628	6701
s9									3089

표 4. 숫자음 "789"의 각 화자별 모델간의 상관계수.

	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	0.81	0.92	0.96	0.90	0.82	0.93	0.59	0.90	0.93
s2		0.77	0.82	0.93	0.85	0.84	0.86	0.76	0.77
s3			0.89	0.85	0.76	0.94	0.56	0.95	0.92
s4				0.92	0.89	0.93	0.61	0.89	0.92
s5					0.91	0.89	0.75	0.83	0.86
s6						0.8	0.69	0.76	0.79
s7							0.65	0.94	0.93
s8								0.57	0.59
s9									0.93

표 5. 숫자음 "789"의 각 화자별 모델간의 비교에 따른 절대차이합.

	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	3334	3592	2459	3171	3587	4361	4886	4834	5134
s2		3538	3114	2050	2948	3767	2617	4570	4758
s3			2189	2563	3013	1667	4670	1886	2471
s4				1985	1940	2475	4333	3011	3128
s5					2139	2746	3566	3413	3512
s6						3287	3639	3705	3728
s7							4750	1636	1954
s8								5285	5442
s9									1745

표 6. 숫자음 "891"의 각 화자별 모델간의 상관계수.

	s2	s3	s4	s5	s6	s7	s8	s9	s0
s1	0.90	0.85	0.95	0.92	0.83	0.86	0.77	0.84	0.71
s2		0.87	0.87	0.90	0.81	0.89	0.83	0.83	0.73
s3			0.86	0.90	0.77	0.91	0.73	0.91	0.77
s4				0.93	0.87	0.85	0.76	0.84	0.70
s5					0.83	0.90	0.77	0.89	0.82
s6						0.72	0.82	0.7	0.65
s7							0.80	0.95	0.81
s8								0.74	0.66
s9									0.85

표 7. 숫자음 "B91"의 각 화자별 모델간의 비교에 따른 절대차이합.

	s2	s3	s4	s5	s6	s7	s8	s9	s10
s1	2329	3846	1857	3030	2809	4816	3660	5303	6762
s2		2967	2412	2315	3002	3674	2687	4247	5700
s3			2735	1995	3219	2000	3290	2238	3943
s4				2228	1835	3696	2935	4095	5537
s5					2722	2583	2930	3043	4165
s6						4038	2666	4515	5588
s7							3476	1498	3191
s8								4052	5283
s9									2727

위의 숫자음별 화자간 비교치의 통계적인 특성을 요약해 보면 표 8과 같다.

표 8. 각 화자별 모델간의 비교에 따른 상관계수와 절대차이합의 통계적 특성. 67BR은 숫자음 "67B"의 상관계수를 67BD는 "67B"의 절대차이합을 나타낸다.

	67BR	67BD	789R	789D	B91R	B91D
최소값	0.38	1877	0.56	1636	0.65	1498
최대값	0.92	6701	0.96	5442	0.95	6762
평균값	0.72	3786	0.83	3302	0.82	3414

일반적으로 숫자음 "67B"이 나머지 두 숫자음보다 더 낮은 상관계수와 더 큰 절대차이합을 보여주었다. 이는 다른 두 숫자음에 없는 이중모음 "ㅝ"이 들어갔기 때문에 화자간의 역동적인 발음차이가 있었기 때문으로 여겨진다. 숫자음 "789"와 "B91"의 음절핵에 해당하는 모음 "아 우 이"는 순서의 차이만 있으므로 거의 비슷한 상관계수와 절대차이합을 나타내어 주고 있다. 화자간의 비교에서도 상관계수가 0.96에 이를 정도로 매우 높은 경우와 절대차이합도 1498에 이를 정도로 유사한 음성으로 분류될 경우가 나타났다. 실제 s7과 s9의 음성을 연구자가 직접 들어봤을 때 거의 같은 화자가 발음한 듯한 청각적 인상을 보여주었다. 기존의 연구에서도 개인별 변화율과 동일한 음성으로 판단하는 지각범위(Yang 2001)를 고려하여 화자확인을 성공적으로 실시하려면 현재의 음향학적인 분석방법에도 여전히 이런 한계점이 있음을 인정하지 않을 수 없다. 이러한 결과는 대역을 다섯 개의 높지 않은 해상도의 간격으로 분류한 것과 관련이 있을 수도 있으므로 앞으로 보다 더 많은 대역을 나눈 경우의 모델생성의 문제점을 찾아볼 계획이다. 아울러, 이전의 연구에서도 보았듯이 단모음의 특성에서 화자간의 차이가 적다면, 여러 개의 정해진 숫자음 가운데 하나를 화자에게 요구하여 발음하게 하여 화자간의 차이를 극대화시키는 방안도 모색되어야 할 것이다(안성주 외 2000). 또한, 저주파 대역에 해당하는 부분을 기준으로 상대적으로 높은 강도를 보이는 지점의 스펙트럼 정보를 이용하여 보다 더 정밀한 화자확인의 문턱값을 지정하는 방안도 시도해볼 수 있을 것이다.

4. 맺음말

이 논문에서는 음향분석의 에러를 최소화할 수 있는 대역별 여과방식을 이용하여 구한 강도값을 구하여 화자내의 변화와 화자간의 비교를 통해 화자확인에 이용할 수 있는 방안을 찾아보았다. 10명의 여성화자가 다섯 번씩 발음한 세 개의 숫자음에 대해 음성부분만을 추출하여 800 Hz 간격의 다섯 개의 대역으로 나누어 여과시킨 뒤 나타나는 각 대역별 강도값을 100등분한 시간점에서 측정한 뒤, 개인별로 비교하여 상관계수와 절대차이합을 구하였다. 이어서, 상관계수가 높고 절대차이합이 적은 비교 쌍의 행렬의 평균을 내어 개인마다 모델로 설정한 뒤, 화자간의 비교를 시도하였다. 그 결과, 화자개인별로 숫자음마다 대체로 상관계수가 높았고, 절대차이합이 적었지만, 변화율이 높은 경우와 낮은 경우로 나눌 수 있었다. 따라서, 동일화자의 발음이라도 모두 화자의 모델설정에 사용하기에는 다른 화자와 동일시될 확률이 높아지게 되므로, 이 연구에서는 상관계수가 높았고, 절대차이합이 적은 쌍을 모델로 설정하여 비교하였다. 화자간 비교에서는 거의 동일한 음성으로 분류될 만큼 높은 상관계수와 낮은 절대차이합을 보이는 비교결과를 보였으므로 앞으로 더 많은 대역을 나누는 문제와 단모음보다는 다수의 이중모음이 포함된 음성자료, 또는 저주파대역에서 높은 강도를 보이는 지점의 좁은대역 스펙트럼 정보와 같은 추가적인 파라미터 이용에 대한 연구가 필요할 것으로 여겨진다.

참 고 문 헌

안성주, 강선미, 고한석, 2000. "가변문턱치와 순차결정법을 통한 문맥요구형 화자확인." *음성과학*, 7(4), 41-47.

양병곤, 2000. "Praat에 의한 숫자음의 음향적 분석법." *음성과학*, 7(2), 127-137.

양병곤, 2001. "남성의 숫자음 발성에 나타난 화자변이." *음성과학*, 8(3), 93-104.

양병곤, 강선미, 2002. "좁은대역 스펙트럼의 차이값과 상관계수에 의한 화자확인 연구." *음성과학*, 9(3), 3-16.

Abercrombie, D. 1967. *Elements of General Phonetics* Chicago, IL: Aldine.

Ladefoged, P. & D. E. Broadbent. 1957. "Information conveyed by vowels." *Journal o f the Acoustical Society of America*, 29, 98-104.

Van Lancker, D., J. Kreiman & K. Emmorey. 1985a. "Familiar voice recognition: Patterns and parameters: Part I. Recognition of backward voices." *Journal of Phonetics*, 13, 19-38.

Van Lancker, D., J. Kreiman & T. D. Wickens. 1985b. "Familiar voice recognition: Patterns and parameters: Part I: Recognition of rate-altered voices." *Journal of Phonetics*, 13, 39-52.

Yang, B. 2001. "Perceptual experiment on number production for speaker identification." *Speech Sciences*, Vol. 8, No. 1, 7-19.

접수일자: 2003. 4. 10.
게재결정: 2003. 5. 29.

▲ 양병곤
부산광역시 부산진구 가야동 산 24 (우: 614-714)
동의대학교 영어영문학과
Tel: +82-51-890-1227
E-mail: bgyang@dongeui.ac.kr
Website: http://www.dongeui.ac.kr/~bgyang