숫자음의 스펙트럼 차이값과 상관계수를 이용한 화자인증 파라미터 연구

A Study on Speaker Identification Parameter Using Difference and Correlation Coeffieicent of Digit_sound Spectrum

이 후 동* · 강 선 미* · 장 문 수** · 양 병 곤***
Hoodong Lee · Sunmee Kang · Moonsoo Chang · Byunggon Yang

ABSTRACT

Speaker identification system basically functions by comparing spectral energy of an individual production model with that of an input signal. This study aimed to develop a new speaker identification system from two parameters from the spectral energy of numeric sounds: difference sum and correlation coefficient. A narrow-band spectrogram yielded more stable spectral energy across time than a wide-band one. In this paper, we collected empirical data from four male speakers and tested the speaker identification system. The subjects produced 18 combinations of three-digit numeric sounds ten times each. Five productions of each three-digit number were statistically averaged to make a model for each speaker. Then, the remaining five productions were tested on the system. Results showed that when the threshold for the absolute difference sum was set to 1200, all the speakers could not pass the system while everybody could pass if set to 2800. The minimum correlation coefficient to allow all to pass was 0.82 while the coefficient of 0.95 rejected all. Thus, both threshold levels can be adjusted to the need of speaker identification system, which is desirable for further study.

Keyword: Speaker Identification, Narrow-Band Spectrogram, spectral energy, difference sum, correlation coefficient

1. 서 론

컴퓨터를 이용하여 사람을 식별하는 방법에는 아이디와 패스워드를 이용하여 인증하는 고전적인 방법을 비롯하여, 사람의 생체정보, 즉 지문이나 홍채패턴과 같은 고정된 신체적 특징이나 음성 및 제스처와 같은 행위적 신체특징을 이용하는 식별 방법이 있다. 특히 사람의 생체정보를 이용하여 인증하는 방법 중 사람의 음성을 이용하여 인증하는 방법은 별도의 고가 장비를 필요로 하지 않으며 인증시 사람이 거부반응을 일으키지 않는다는 장점이 있다.

* 서경대학교 컴퓨터학과
** 서경대학교 소프트웨어학과
*** 동의대학교 영어영문학과

일반적으로 화자인증은 사전에 구축된 개별화자 모델과 인증을 위해 입력된 음성과의 유사도를 측정하여 일치여부를 확인하게 된다. 여기에 사용되는 파라미터들로는 MFCC(Mel Frequency Cepstral Coefficient)와 음성의 음향학적 특징인 포만트, 피치, 강세 등의 정보를 사용한다[3][4]. 그러나 포만트, 피치 등의 정보들이 개별화자의 특징을 잘 반영하는 반면, 음성으로부터 추출되는 정보가 부정확하다는 문제점을 가지고 있다. 이러한 문제점을 개선하기 위하여 본 논문에서는 광대역 스펙트럼에 비하여 협대역 스펙트럼이 비교적 안정적인 값을 추출할 수 있다는 것에 착안하여, 협대역 스펙트럼의 정보가 화자인증 파라미터로 유효한지에 대해서 검증하고자 한다.

2장에서는 본 논문의 연구배경과 선행연구에 대해서 설명하며, 3장에서는 본 논문에서 제안하는 화자인증 방안에 대해서 기술한다. 4장에서는 제안하는 화자인증 방안을 실험을 통하여 검증하고, 5장에서 결론 및 향후 연구계획에 대해서 기술한다.

2. 연구배경 및 선행연구

2.1 연구배경

기존의 화자인증 파라미터 연구로 음성의 음향학적 특징은 포만트, 피치 등의 정보를 사용하였다. 그러나 이러한 음향학적인 특징의 추출에 있어서 조음기관은 매우 유연하고 연속된 움직임을 보이지만 음향학적 측정값은 발성기관의 자연스럽고 느린 변화라는 기본 가정에 어긋나는 급작스런 오류값들이 많이 나타난다. 오류가 많은 음향학적 특징값을 근거로 생성된 화자별 모델을 통해서 화자인증을 한다는 것은 화자인증 시스템의 성능에 좋지 않은 영향을 끼치게 된다. 그에 반해 푸리에 변환을 통해 구해지는 스펙트럼의 정보는 정확한 측정이 가능하다. 스펙트럼은 분석구간의 길이에 따라 광대역 스펙트럼과 협대역 스펙트럼으로 구분된다. 그림 1은 여성화자가 발음한 모음 '아'에 대한 광대역 스펙트럼과 협대역 스펙트럼의 일부를 보여주고 있다.

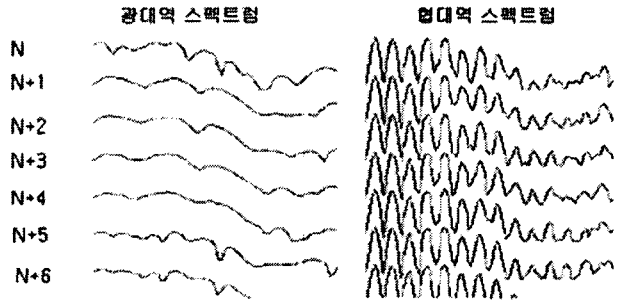

그림 1. 광대역 스펙트럼과 협대역 스펙트럼

광대역 스펙트럼의 경우 많은 구간에서 이웃하는 구간의 스펙트럼과 많은 차이를 보이고 있다. 반면에 협대역 스펙트럼의 경우 다소 안정적이다. 고주파 영역으로 갈수록 약간의 차이를 보이고 있으나 광대역에서 발견되는 급작스런 변화는 보이지 않는다[3]. 따라서 본 논문에서는 동일 음성

내에서도 많은 차이를 보이는 광대역 스펙트럼이 아닌, 변화가 적고 안정적인 협대역 스펙트럼을 사용하여 음성 모델을 생성한다.

2.2 선행연구

피치나 포만트를 사용하여 화자간의 구별을 하기에는 그 값들의 추출의 정확성에 문제가 있을 뿐만 아니라 단순히 그 값들로 화자별 모델을 생성하기에는 단순하다는 문제점이 있다. 그래서 피치나 포만트보다 정확하게 구할 수 있으며 더 많은 정보를 담고 있는 스펙트럼의 정보를 이용하여 화자별 모델을 생성하고 화자인증에 사용한다. 협대역 스펙트럼 정보의 화자별 변별력 실험을 위해 숫자음 스펙트럼 모델을 생성한다. 숫자음 모델과의 비교는 두 스페트럼 강도의 차이의 절대값의 합(이후 절대차이합으로 표현)과 스펙트럼의 전체적인 모양의 유사성을 비교하는 상관계수를 사용하여 비교한다.[1] 절대차이합은 식 1을 이용하여 계산한다. 여기서 Ha와 Hb는 비교되는 음성에서의 스펙트럼의 강도를 나타낸다.

$$DifferenceSum = \sum_{n=1}^{N} |Ha[n] - Hb[n]| \tag{1}$$

상관계수는 식 2와 같이 두 변수의 값이 연속적 측정값으로 주어지는 경우에 적용되는 피어슨 상관계수[7]를 사용한다.

$$(H_{a_1}, H_{b_1}), \ldots\ldots\ldots (H_{a_n}, H_{b_n})$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

$$r_p = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}$$

스펙트럼의 절대차이합만을 사용하여 비교할 경우 강하게 말할 때와 약하게 말할 때 차이가 있으므로 스펙트럼의 전체적인 모양을 비교하는 상관계수를 구하여 높은 상관계수를 보이는 경우 비슷한 스펙트럼을 보이는 것으로 간주하여 동일화자로 처리한다.

3. 제안하는 화자인증 방법

3.1 협대역 스펙트럼을 이용한 숫자음 모델 생성

기존의 연구[1]에서는 하나의 숫자음을 대상으로 실험을 하였고, 화자인증용으로 사용될 수 있는 가능성을 확인하였다. 그러나 하나의 음만으로는 절대적인 비교 데이터량이 부족할 뿐만 아니

라, 여러 모음에서 나타날 수 있는 다양한 음성적 특징을 비교할 수 없는 문제가 있다. 따라서 본 논문에서는 서로 다른 모음을 포함하는 3 개의 음으로 이루어진 숫자음을 사용하여 숫자음 모델을 생성한다. 실험에 사용된 숫자음은 표 1과 같다.

표 1. 실험에 사용한 단어들

첫째음절	둘째음절	셋째음절	녹음단어
일 (1)	삼 (3)	육 (6)	'일삼육', '일삼영' '일사육', '일사영' '일팔육', '일팔영'
이 (2)	사 (4)		'이삼육', '이삼영' '이사육', '이사영' '이팔육', '이팔영'
칠 (7)	팔 (8)	영 (0)	'칠삼육', '칠삼영' '칠사육', '칠사영' '칠팔육', '칠팔영'

실험데이터는 이, 아 ,유, 여가 섞여 있는 숫자음으로 한다. 모음 '우'로 된 5, 9는 저주파 지역에 데이터가 밀집되어 있어서 실험데이터에서 제외한다. 5, 9를 제외한 숫자를 이용하여 총 18 가지의 세음절로 구성된 숫자음 조합 단어를 사용한다.

3.2 배음 스펙트럼

세 개의 음절로 구성된 데이터에서 각각의 음절구간을 검출하기 위하여 다음과 같은 방법을 사용한다.

1) 음성의 강도곡선(Intensity Contour)을 계산한다.
2) 음성의 강도곡선에서 전체평균을 구하여 평균 + (표준편차 / 2)값을 근거로 하여 세 개의 지역을 구한다.
3) 구해진 세 개 지점에서 피치가 0이 아닌 구간에서부터 최고 정점 구간을 계산하여 배음 스펙트럼의 추출의 시작값으로 사용하며 그 결과는 그림 2와 같다.

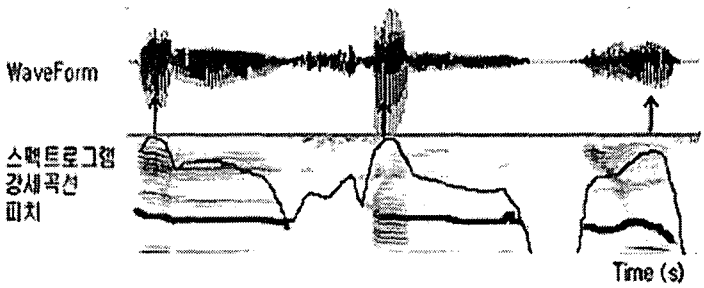

그림 2. 음절 구간의 검출

위와 같은 방법으로 얻어진 세 개의 음절구간 시작 시간을 기점으로하여 7 개의 단구간 스펙트럼을 구한다. 스펙트럼을 구하기 위해 20 ms의 Hamming window를 사용하며, 그림 2에서와 같이 5 ms씩 이동함으로서 숫자음 내에서 자음과 잡음 구간의 영향을 받지 않는 모음구간에 대한 특징만을 살핀다. 모음마다 내재적 길이가 다르기는 하지만 35 ms의 구간이면 충분히 모음구간만을 고려할 수 있다. 개별 화자의 음성의 특징을 잘 살펴보기 위해 고주파대역강조(preemphasis)를 적용한다. 구해진 스펙트럼으로부터 남성화자의 피치에 가까운 임의의 간격인 150 Hz를 기준으로 하여 1,000 Hz 단위로 4 개씩의 배음 스펙트럼, 총 20 개의 배음 스펙트럼의 강도값을 추출한다.

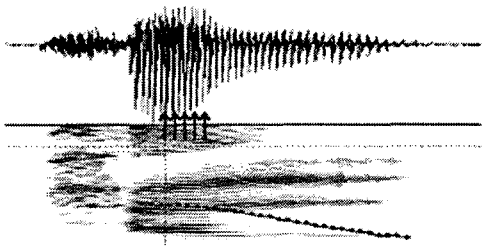

그림 3. 협대역 스펙트럼의 측정

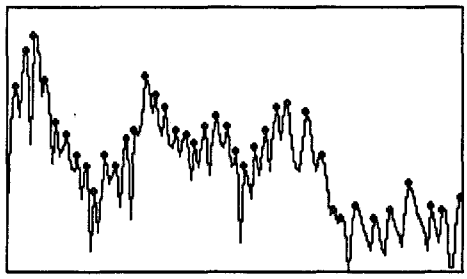

그림 4. 배음 스펙트럼의 추출

3.3 숫자음 모델의 생성

숫자음 모델은 앞 절에서 설명한 배음 스펙트럼의 강도값을 이용하며, 화자별로 18개의 숫자음에 대해 각각 모델을 생성한다. 숫자음의 기본 모델은 식 3과 같이 계산한다.

$$\text{숫자음모델} = \frac{1}{N}\sum_{i=1}^{N}(H_i) \qquad (3)$$

H_i: 배음스펙트럼의강도값

N: 녹음단어수 * 음절구간 * 단구간수 * 배음스펙트럼수

하나의 녹음단어에는 세 개의 음절구간이 추출되고, 한 음절구간에서는 앞 절에서 언급한 것처럼 7 개의 단구간이 사용된다. 그리고 각 단구간에서 20 개의 배음 스펙트럼 강도값이 추출되므로, 여기서 추출되는 모든 강도값의 평균으로 숫자음 모델을 만든다. 그리고 녹음오류를 감안하여 식 3으로 도출한 평균값과 차이가 큰 녹음 문장을 제거하고 나머지 녹음 단어만으로 평균값을 다시 계산하여 최종 숫자음 모델을 생성한다. 그림 5는 녹음 단어가 5 개일 때 최종 숫자음 모델을 생성하는 과정을 나타내고 있다.

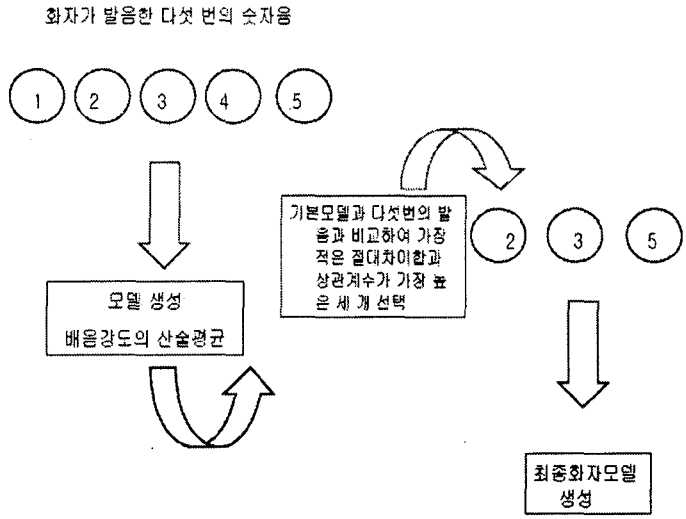

그림 5. 최종 숫자음 모델 생성

4. 실험 및 검증

4.1 음성데이타 수집

실험에 사용된 데이터는 서울에 거주하는 20대 중반 남성화자 4 명을 대상으로 수집하였다. 실험에 참가한 화자는 표 2와 같다. 녹음은 조용한 연구실에서 Pentium 4 데스크탑 컴퓨터에서 SHURE사의 SM58S 마이크를 통해 입력하였으며, 표 2에 나와 있는 녹음단어별로 10 회씩 반복하여 녹음한다. 정확한 실험데이터 수집을 위해 직접 제작한 음성 분석 소프트웨어[6]를 사용한다. 본 음성분석 소프트웨어에서는 세 개의 음절로 이루어진 숫자음 녹음시 검출된 음절구간을 피험자에게 제시함으로써 피험자는 잘못 검출된 부분에 대해서는 재녹음을 통해 수집할 수 있도록 한다. 녹음된 자료는 11,025 Hz 샘플링률과 16 bit로 양자화한다. 총 10 회 수집한 데이터 중에서 5 회는 숫자음 모델을 만드는데 사용하며, 나머지 데이터는 인증 실험에 사용한다.

표 2. 실험에 참가한 화자정보

화자	거주지	나이	키(cm)
s1	서울	28	178
s2	경기	26	174
s3	경기	24	174
s4	경기	27	180

4.2 숫자음 모델간의 비교

여기에서는 숫자음 모델간의 절대 차이합과 상관계수가 어느 정도의 값으로 나오는지를 확인하며, 숫자음중에서 화자별 특징을 잘 보이는 숫자음을 추출하여 다음 인증실험에 이용하도록 한다.

화자간 숫자음 모델의 값을 비교하는 실험을 통해 숫자음 중에서 화자별로 큰 절대차이값을 갖는, 그리고 낮은 상관계수값을 갖는, 즉 화자인증시 좋은 결과를 보일 수 있는 숫자음을 찾는다. 18개의 숫자음 모델간의 비교를 통해 구해진 절대차이합의 평균과 표준편차를 그림 6에 나타내었으며, 그림 7은 상관계수의 평균과 표준편차를 나타낸다. 절대차이합의 평균이 크고, 상관계수의 평균이 낮은 숫자음 모델일수록 화자간의 특징을 잘 보여주는 숫자이다. 여기서 선택된 숫자음을 화자인증시 사용한다면 더 좋은 성능을 보일 것이다.

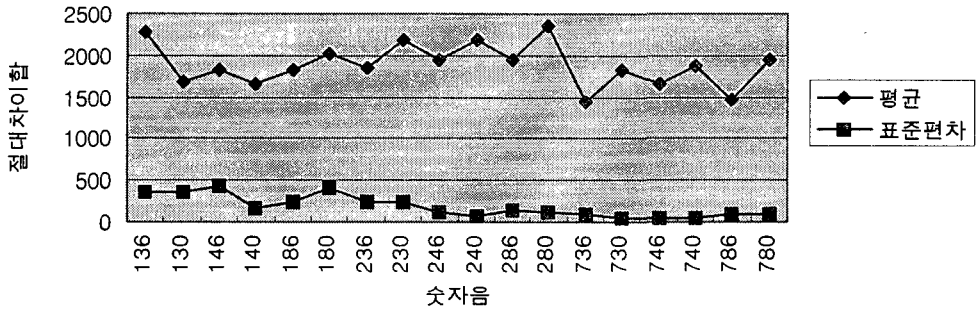

그림 6. 절대차이합의 평균과 표준편차

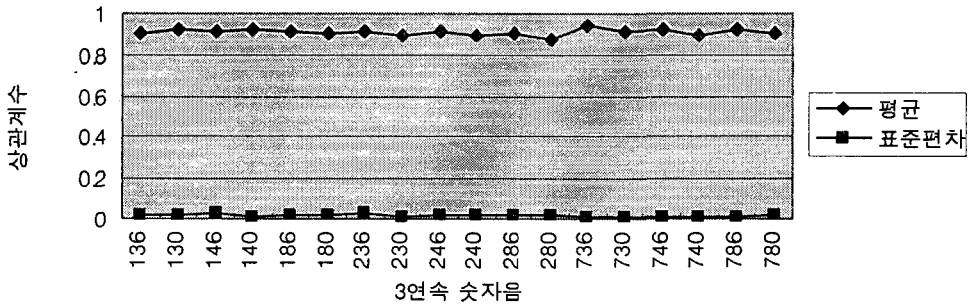

그림 7. 상관계수의 평균과 표준편차

최적 숫자음모델을 추출하기 위하여 식 4를 사용한다. 식 4는 평균값과 표준편차의 가중치를 2:1로 적용한 경우를 나타낸다. 여기서 S_a는 절대차이합의 평균, S_d는 표준편차를 나타내며, C_a는 상관계수의 평균, 그리고 C_d는 표준편차를 나타낸다.

$$R=(S_a-\frac{d_s}{2}(S_d+\overline{S_d}))-(C_a-\frac{d_c}{2}(C_d-\overline{C_d})) \quad (4)$$
$$d_s=\frac{S_a}{S_d},\quad d_c=\frac{C_a}{C_d}$$

18 개의 숫자음을 위의 식 4를 이용하여 계산한 후, 가장 큰 값을 갖는 4 개의 숫자음을 선택한다. 본 실험에서 사용된 데이터 범위에서는 '246', '240', '286', '280'의 숫자음 모델이 선택되었다.

4.3 화자간의 비교

앞 절에서 숫자음 모델간의 비교를 통해 화자인증시 좋은 성능을 가질 것으로 판단되는 숫자음 4 개를 선정하였다. 선정된 4 개의 숫자음모델에 대해서 동일화자의 데이터와 다른 화자의 데이터를 비교함으로써, 제안한 파라미터가 화자간 변별력을 가지는지 살펴본다. 실험을 위해 화자별로 숫자음 모델과 모델 생성시 사용한 5 개의 녹음단어를 제외한 5 개 녹음문장 데이터를 이용하여 절대차이값과 상관계수값을 구한다. 실험을 통해 스펙트럼 정보가 화자별로 구분 지을 수 있는 특징으로 사용될 수 있는지를 확인한다.

그림 8과 그림 9에서 동일 화자의 개별 숫자음 데이터와 다른 화자의 개별 숫자음 데이터와의 비교 결과를 보인다. 그림 7에서 보듯이 s3 화자를 제외한 나머지 화자에서 다른 화자의 개별 숫자음 데이터와의 절대차이합이 2000 이상의 차이를 보이며, 동일 화자의 숫자음 데이터와는 1500 정도의 값을 보인다. 그림 8에서도 s3를 제외한 나머지 화자의 상관계수값이 동일 화자내에서는 0.92 이상의 값을 보이며, 다른 화자와는 0.87 이하의 값을 갖는다. 따라서, 동일화자의 데이터와 다른 화자의 데이터가 명백하게 구분됨을 알 수 있다. s3 화자의 경우 s3 개별 데이터와 다른 화자 개별 데이터의 비교값에서 다른 화자에 비해서 작게 나왔다. 이것은 다른 s3 화자의 음성이 다른 화자와 비슷한 발성을 보이거나, 음성데이터 수집시 안정된 발성을 못한 것으로 인해 나온 결과로 판단된다. s3화자에 대해 계속적으로 음성데이터를 수집하고, 실험을 실시하여 그 원인을 밝혀낼 필요가 있다.

여기서는 화자별 개별 데이터를 이용하여 다른 화자와의 차이와, 화자 내에서의 차이를 살펴보았다. 지금까지 살펴본 것처럼 스펙트럼 정보가 화자별로 변별력을 갖는다는 것을 확인할 수 있다.

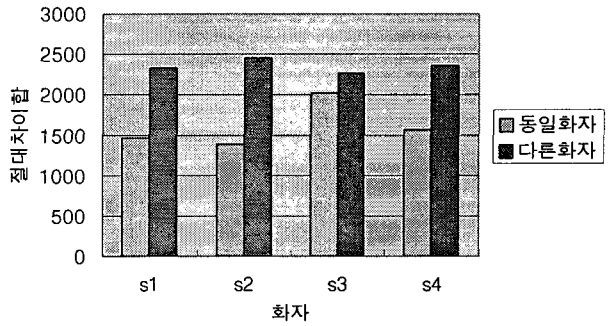

그림 8. 숫자음 모델과 개별 숫자음 비교 : 절대차이합

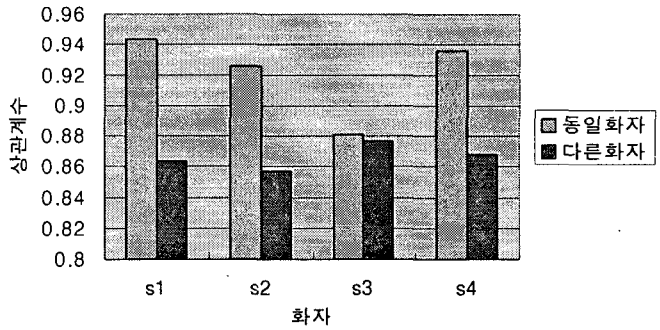

그림 9. 숫자음 모델과 개별 숫자음 비교 : 상관계수

4.4 화자인증 실험

여기서는 절대차이합과 상관계수의 임계값을 변경하면서 화자인증 실험을 실시한다. 그림 10은 절대차이합의 임계값에 대해서 인증률과 거절률을 보이며, 그림 11은 상관계수의 임계값에 대한 결과이다.

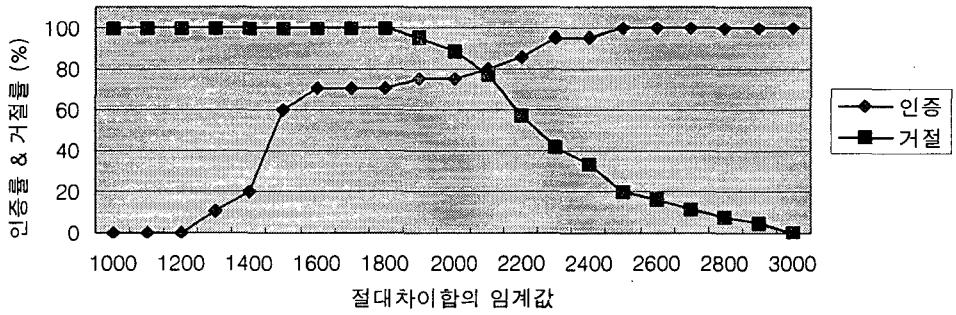

그림 10. 절대차이합 임계값의 변화에 따른 인증률과 거절률의 관계

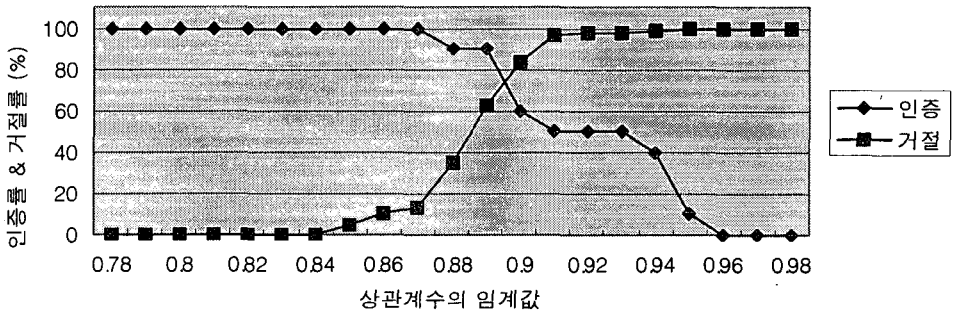

그림 11. 상관계수 임계값의 변화에 따른 인증률과 거절률의 관계

절대차이합의 경우 임계값을 크게 조절하면, 인증률은 상승하는 반면 거절률은 낮아지게 되며, 상관계수의 경우는 임계값을 크게 할수록 거절률은 상승하는 반면 거절률은 낮아지게 된다. 임계값은 대개 인증률과 거절률이 같은 값을 갖는 지점을 사용한다. 절대차이합만을 고려한 경우 2100값을 갖으며, 상관계수의 경우 0.9의 값을 갖는다.

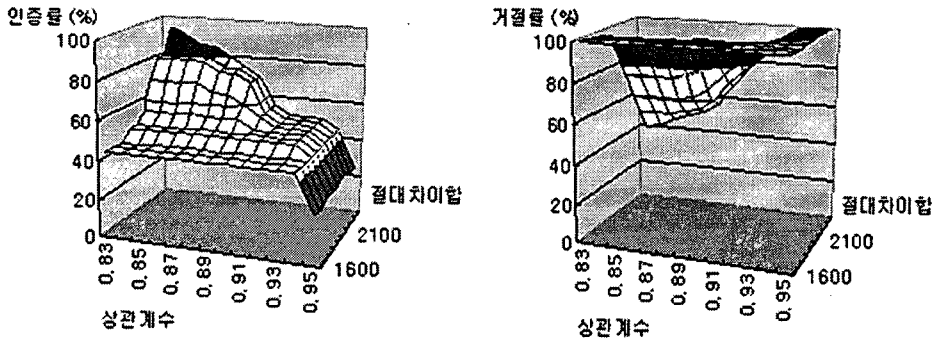

그림 12. 절대차이합과 상관계수의 임계값의 변화에 따른 인증률과 거절률의 관계

그림 12는 화자인증시 절대차이합과 상관계수에 대한 인증률과 거절률을 보이고 있다. 이 경우 절대차이합이 임계값 이하이면서 상관계수가 임계값보다 높은 경우 인증을 하며, 그 둘중에 하나라도 만족되지 않을 경우 거절하게 된다. 예를 들어, 그림 11에서처럼 상관계수의 임계값이 0.87인 경우 거절률이 20% 이하인 반면에, 절대차이합과 상관계수의 임계값을 적용하였을 시에는 거절률이 50% 이상으로 높아진다. 따라서 절대차이합이나 상관계수 단독으로 적용했을 때보다 화자인증시 다른 화자에 대한 거절 기능이 좋아진다.

절대차이합과 상관계수의 임계값을 결정하는 것이 화자인증 시스템의 성능을 결정하는 중요한 문제이다. 화자인증 시스템이 사용되는 분야의 특징에 따라서 알맞은 절대차이합과 상관계수의 임계값을 설정한다면 화자인증시 목적에 부합하는 성능을 얻을 수 있다.

5. 결론 및 향후계획

본 논문에서는 화자음성의 협대역 스펙트럼이 시간축에 대해 비교적 변화가 적고 안정적인 값을 나타낸다는 점에 착안하여 화자음성의 스펙트럼 정보를 이용한 화자인증 방안을 제안하였다. 스펙트럼 강도값에 대한 절대차이합과 스펙트럼 분포에 대한 상관계수를 이용하여 화자별 음성모델을 생성하고, 생성된 모델과 녹음된 테스트 음성을 비교하여 화자별 변별력을 살펴보았다.

실험모델의 최적화를 위하여 18 개 숫자음 모델에 대해서 실험을 통하여 최적의 숫자음 모델을 추출하여 화자인증용 파라미터로 사용하였다. 그 결과, 본 논문에서 사용한 실험용 음성데이터에 대해서 화자간 변별력이 충분히 존재함을 인증실험을 통하여 확인하였다.

제안하는 화자확인 파라미터를 보다 일반화시키기 위해서는 다양한 화자에 대한 변별력 실험을 수행할 필요가 있으며, 이를 통하여 화자간의 특징적 차이를 찾는 연구를 계속 할 필요가 있다. 또한, 본 논문의 실험에서도 나타난 것처럼 화자에 따라서는 화자내에서의 발성의 변이가 큰 경우가 나타나므로, 이것을 수용할 수 있는 음성모델을 생성하는 연구가 필요하다.

참 고 문 헌

[1] 양병곤. 2002. "좁은대역 스펙트럼의 차이값과 상관계수에 의한 화자확인 연구." *음성과학*, 9권 3호.
[2] 양병곤. 2002. "남성의 숫자음 발성에 나타난 화자변이." *음성과학*, 8권 3호, 93-104.
[3] 강선미 외. 2001. "화자인식을 위한 화자 고유의 음성특징추출과 적응모델에 관한 연구 2차년도." 과학재단 특정기초과제 연구보고서.
[4] 강선미 외. 2002. "화자인식을 위한 화자 고유의 음성특징추출과 적응모델에 관한 연구 3차년도." 과학재단 특정기초과제 연구보고서.
[5] 구희산, 고도흥, 양병곤, 김기호, 안상철. 1988. *음성학과 음운론*, 서울: 한신.
[6] 이후동, 강선미, 장문수. 2004. "사용자 편의성을 고려한 음성분석 소프트웨어의 구현." 통신학회 하계학술대회 논문집.
[7] 류근관. 2003. *통계학*, 법문사.

접수일자: 2004. 07. 30
게재결정: 2004. 08. 31

▲ 이후동
서울특별시 성북구 정릉 4동 (우: 136-704)
서경대학교 컴퓨터과학과
Tel: +82-2-940-7291, Fax: +82-2-919-5075
E-mail: hdlee77@empal.com

▲ 강선미
서울특별시 성북구 정릉 4동 (우: 136-704)
서경대학교 컴퓨터과학과
Tel: +82-2-940-7291, Fax: +82-2-919-5075
E-mail: smkang@skuniv.ac.kr

▲ 장문수
서울특별시 성북구 정릉 4동 (우: 136-704)
서경대학교 소프트웨어학과
Tel: +82-2-940-7509, Fax: +82-2-919-5075
E-mail: cosmos@skuniv.ac.kr

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