영상정보 보완에 의한 음성인식
(Speech Recognition with Image Information)

李 天 雨*, 李 相 杭**, 梁 根 模***, 朴 仁 政****
(Chun-Woo Lee, Sang-Won Lee, Keuhn-Mo Yang, In-Jung Park)

요 약

음성의 인식율 저하는 주로 잡음에 의해 발생하고, 이러한 요인을 제거하기 위해 주로 필터뱅크를 사용하여 왔지만, 본 논문은 2 차원 선형예측이라는 영상 특징 추출 방법을 이용하여 잡음에 강인한 숫자음 인식을 시도하였다. 먼저, 음성에 대한 인식결과를 도출하기 위해, 13 차 선형예측 계수를 이용하여 인식을 시도하였다. 이 때, 잡음을 추가한 음성을 이용하여 시험한 결과, 5 개의 숫자음, '영', '사', '오', '육', '구 '에서 인식결과의 저하를 볼 수 있었다. 이러한 결과를 향상시키기 위해 2 차원 선형예측 계수를 추가한 인식기 입력 데이터를 구현하였다. 이 때, 선형예측 계수는 각 프레임별로 추출하였고, 음성 데이터와 합한 영상 데이터를 가지고 인식 실험을 실시하였다. 이 때, 숫자음 '사' 와 '구' 에 대해서는 상당한 향상을 보였다.

Abstract

The main factor decreasing speech recognition rate is the surrounding noise. To lower the noise effect, we generally used the filter bank at preprocessing stage. But, in this paper, we tried to recognize the 10 numeral numbers using 2-D LPC to extract image feature. At first, we obtained the result of speech-only recognition using 13th-order LPC coefficients and then, for distorted speech recognition results of '0','4','5','6','and 9', we added image parameters such as 12th-order 2-D LPC coefficients. At each frame, we extracted the 2-D LPC coefficients, and simulated recognizer with two parameters such as speech and image. Finally, for the numbers, such as '4' and '9', the better results were obtained.

I. 서 론

미디어는 3 가지 즉, 문자 정보, 소리정보, 이미지 정보로 분류할 수 있는데, 이러한 미디어들의 상호관계를 보면, 문자와 소리사이에는 음성인식과 TTS(Text-To-Speech), 문자와 비디오 사이에는 문자인식 및 입술 읽기, 소리와 비디오 사이에는 입술동기화, 얼굴 애니메이션 그리고 결합 오디오/비디오 코딩과 관계가 있다[1].

기존의 음성 인식 방법은 음성 그 자체만을 가지고 특징을 구한 후, 바로 인식을 시도하였지만, 본 논문에서는 소리와 이미지와의 상호관계를 이

* 正會員, 又松 情報大學 電算情報系列
(email : cwlee@zerobit.woosonginfo.ac.kr)
** 正會員, 檀國大學校 電子工學科
(email : sephia@shinbiro.com)
*** 準會員, 檀國大學校 電子工學科
(email : hanair@anseo.dankook.ac.kr)
**** 正會員, 檀國大學校 電子工學科
(email : ijp48128@anseo.dankook.ac.kr)

용하여 음성 인식을 시도하고자 한다.

먼저, 음성과 영상은 인식기에 알맞는 형태로 특징을 추출하여야 하는데, 그 방법으로는 성도 모델을 기본으로 하는 LPC, 공진 주파수를 이용하는 포만트, LPC로부터 유도되는 켑스트럼 등과 청각 모델을 기본으로 하는 주파수 필터 뱅크를 이용하는 방법이 있다.

이러한 특징들은 인식을 수행하는 인식기에 입력되고, 이러한 입력기들은 DTW, 신경회로망, HMM 등 과 같은 것들을 사용하는데, 음성 자체만으로 인식을 시도함으로, 환경적인 요인에 의해 인식결과에 많은 영향을 받는다. 음성 인식 시스템의 성능에 영향을 끼치는 외부 요인으로는 동일 화자에 대해서도 나타나는 입력레벨 변화, 신호와 상관성이없는 부가적인 정지성 교란신호에 의한 부가적인 잡음, 음성이 선형 시불변 필터를 통과할 때 나타나는 스펙트럼의 왜곡, 성도의 크기나 모양에 의해 나타나는 생리학적인 차이 그리고 cocktail 파티 효과와 같은 다른 화자에 의한 간섭 등이 있다[2].

이처럼, 음성 한가지로만 인식을 수행할 경우 여러 가지 외부 요인에 의해 인식율이 저하되는데, 이렇게 저하되는 인식율을 높여 보고자, 본 논문에서는 음향적인 잡음에 영향을 받지 않고 음성인식에 기여할 수 있는 가시적 데이터와 함께 인식을 시도할 것이다. 가시적인 데이터는 배경 잡음에 상관없이 항상 음성 정보를 담고 있으며, 따라서, 그러한 정보를 이용하여 인식을 시도하면 더 좋은 인식율을 보일 수 있을 것으로 본다.

본 논문에서는 음성 데이터와 영상 데이터에 대하여 선형예측 방법을 사용한다. 이는 음성에서는 현재의 샘플값이 과거의 샘플값들에 의해 추정될 수 있다는 것을 의미하고, 영상에서는 현재 샘플의 농도값이 이전의 농도값들에 의해 추정될 수 있다는 것을 의미한다. 이러한 방법은 음성과 영상이 서로 동일한 방법을 사용했다는 점에서 시스템의 부피도 줄일 수 있고 또한, 이렇게 얻어진 데이터는 음성인식에 상당한 도움이 될 수 있다는 것을 알 수 있다.

음성에서의 선형예측 계수들과 영상에서의 선형예측 계수들은 다층 신경회로망의 학습 계수로써 사용되고, 교사신호를 통해 최적화 된다. 결과적으로 얻어지는 가중치 벡터들은 음성 인식기 실험에 사용된다.

이 때, 실험은 가우션 잡음 생성기를 이용하여 시험에 사용되는 음성에 균일하지 않은 데이터를 추가하며, 각각의 음성에 대해 선형예측 계수를 추출한 후, 왜곡된 부분에 대해 영상정보를 이용하여 향상된 결과를 제시한다. 여기서 히스토그램을 이용한 경우와 비교를 들어 선형예측 계수의 효율성을 나타낸다[11].

Ⅱ. 2 차원 선형 예측법

카메라 입력 신호는 식 (1) 에 의해 그레이 레벨로 변환이 이루어지고, 각각의 영상 프레임들은 선형 예측을 위해 사용된다.

$$Y = 0.299000R + 0.58700G + 0.11400B \quad (1)$$

(a) 숫자음 '일'

(b)숫자음 '오'

(c) 숫자음 '구'

그림 1. 숫자음 '일', '오', '구' 의 그레이 레벨 변환 결과

그림 1 은 카메라 입력신호를 그레이 레벨로 변환한 결과를 보이고 있다. 그림 1 을 보면, 주위의 농도값의 변화가 급격하게 이루어지지 않고, 서로 연관되어 변화하는 것처럼 보인다. 가로 방향과 세로 방향에 대한 변화가 급격하지 않은 것은 음성에서의 신호 파형이 갑작스럽게 변화하는 것이 아니라, 서로 상관성을 갖고 변화한다는 것으로 생각할 수 있다. 이러한 생각을 바탕으로 그림 2 를 보면, 가운데 위치한 $g(i,j)$ 를 예측할 수 있는데 식 (2) 에 나타나 있다.

$$g(i,j) = a_1 \cdot g(i-1,j-1) + a_2 \cdot g(i-1,j) + \cdots + a_8 \cdot g(i+1,j+1) \quad (2)$$

여기서, $g(i,j)$ 는 i 와 j 위치의 농도를 나타내고, $a_1, a_2, \cdots, a_8$ 은 예측계수들을 나타낸다.

$\rightarrow i$

$g(i-1,j-1)$	$g(i,j-1)$	$g(i+1,j-1)$
$g(i-1,j)$	$g(i,j)$	$g(i+1,j)$
$g(i-1,j+1)$	$g(i,j+1)$	$g(i+1,j+1)$

그림 2. $g(i,j)$ 를 중심으로 한 각 화소 위치

영상은 주로 지그재그 형식으로 보여지는 라스타 주사 데이터에 의해 이루어지므로, 식 (2)를 보다 예측가능한 방식으로 표현하면, 식 (3) 과 같다.

$$\hat{g}(i,j) = a_1 \cdot g(i-1,j-1) + a_2 \cdot g(i,j-1) + a_3 \cdot g(i-1,j) \quad (3)$$

이것은 주위 화소에 의한 것보다도 주사방식에 의해 만들어진 예측 방법이다. 실제 이 식에서 필요한 것은 주위의 화소 값이 아닌, 예측을 하기 위한 예측계수들이므로 상관관계식을 유도해보면, 식 (4) 와 같다.

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_{p-1} \\ R_1 & R_0 & \cdots & R_{p-2} \\ & \vdots & & \\ R_{p-1} & R_{p-2} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ R_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_p \end{bmatrix} \quad (4)$$

여기서, 예측 계수들은 Levinson-Durbin 알고리즘을 이용하여 구하게 되고, 결과적으로 생성되는 a_1, a_2, .., a_p 값들이 구하고자 하는 예측 계수들이 된다.

식 (4)에서 각각의 자기상관 함수들은 다음과 같은 의미를 포함하고 있다.

R_0 는 전체 영상 영역의 2 제곱 평균이고, R_1 은 원 영상과 원 영상을 왼쪽으로 1 위로 1 이동한 영상과의 2 제곱 평균, R_2 는 원 영상과 원 영상을 위로 1 화소 이동한 영상과의 2 제곱 평균, 그리고 R_3 은 원 영상과 원 영상을 왼쪽으로 1 화소 이동한 2 제곱 평균이다.

III. 인식 실험

인식에 사용된 음성데이타는 샘플링 주파수가 8 kHz 이고, Levinson-Durbin 알고리즘을 이용하여 13 차 계수를 구하였다[5][6][7]. 각 음성 프레임은 음성정보의 손실을 줄이기 위해 1/2 중첩하여 사용하였으며 각 숫자음별 계수를 구하였다. 또한, 위에서 얻은 입술정보에 대한 데이터들은 자기 상관 함수를 통한 12 차 선형예측 계수를 구하였는데, 숫자음 '6' 의 총 프레임인 6 프레임을 기준으로 각 숫자 음에 대하여 선형예측 계수들을 구하였다. 이러한 특징들은 신경회로망의 입력으로 사용되기 위해, 그림 3 과 같이 영상 데이터 와 음성 데이터를 서로 합하여 사용되었다.

영상 데이타	음성 데이타
입술에 대한 12 차 6 프레임 LPC 계수	13 차 15 프레임 LPC 계수

그림 3. 입력으로 사용되는 패턴 합성 방법

사용된 음성은 단일 화자의 음성을 사용하였고, 원 음성에 잡음을 부가하기 위해 가우션 잡음 생성기를 사용하였다. 각 잡음 레벨에 따라 각기 다른 선형예측 계수를 구하였으며, 영상 데이타를 부가한 실험을 실시하였다.

학습을 위해 사용된 데이터는 0 dB 음성 데이터이며, 잡음이 추가된 음성은 시험을 위해 사용되었다. 또한 사용한 다층 신경회로망은 입력 노드 수가 285 개, 은닉층 노드 수가 8 개, 출력층 노드 수가 10 개인 신경회로망이고, 개선된 역전파 알고리즘을 사용한 학습방법을 사용하였다.

Ⅳ. 검토 및 결론

위 실험은 잡음이 존재한다는 상황을 가정하기 위하여 가우션 잡음 발생기를 사용하였으며, 각 잡음 레벨별로 실험을 실시하였다. 이 때 음성정보에 영상정보를 추가하는 것이 음성적 왜곡이 발생한 숫자음에 대하여 보다 효과적인 결과를 보였다.

표 1 의 음성만으로 입력시, 숫자음 '영', '사', '오', '육', '구' 만이 잡음 레벨별로 상당히 왜곡되어 출력됨을 볼 수 있다. 특히, 숫자음 '오' 는 20 dB 잡음 추가시 급격히 출력결과가 줄어들기 시작하였고, 다른 숫자음들의 출력 결과들도 계속적으로 줄어듬을 알 수 있었다. 이렇게, 출력결과가 저조한 음성에 대해서 2 차원 선형예측 계수를 보완한 경우의 결과를 표 2 에 보였다.

표 2 에서 보면, 숫자음 '사', '구' 의 출력 결과가 상당히 증가한 반면, 숫자음 '영', '오', '육' 은 약간 증가하거나 저조한 결과를 나타내고 있음을 알 수 있다. 히스토그램을 보완한 결과와 비교한 경우, 숫자음 '육' 에서만 차이가 날 뿐 히스토그램보다 더 효과적인 결과를 얻을 수 있었다. 이것은 음성 자체의 정보 보다는 영상정보를 추가한 것이 좀더 나은 결과를 보일 수 있다는 것을 보이고, 히스토그램이라는 영상정보 보다는 2 차원적인 선형적인 관계를 유도한 선형예측 계수를 사용한 것이 좀더 효율적이고 효과적이라는 것을 보여주고 있다. 결과적으로, 영상정보라는 것은 비 선형적인 신호에 의해 왜곡된 음성정보를 보조해주며, 적은 양의 정보만으로도 인식결과를 향상시킬 수 있다는 것을 보여준다.

참고문헌

[1] TSUHAN CHEN, MEMBER, IEEE, AND RAM R.RAO, "Audio-Visual Integration in Multimodal Communication", Proceedings of the IEEE, VOL. 86, NO. 5, pp.837~852, MAY 1998.

[2] Alejandro Acero, "Acoustical and Envinonmental Robustness in Automatic Speech recognition", Kluwer Academic Publishers, 1993.

[3] RANDY CRANE, "A Simplifed Approach to Image Processing", Prentice-Hall, 1997.

[4] Ioannis Pitas, "Digital Image Processing Algorithms" Prentice Hall, 1993.

[5] John R. Deller, Jr., John G. Proakis and John H. L. Hanson,"Discrete-Time Processing of Speech Signal", Macmillian Publishing Company, 1993.

[6] Lawrence Rabiner, Biing-Hwang Juang, "FUNDAMENTAL OF SPEECH RECOGNITION", Prentice-Hall International, Inc. 1993.

[7] Allen Gersho, Robert M. Gray, "VECTOR QUANTIZATION AND SIGNAL COMPRESSION",

KLUWER ACADEMIC PUBLISHERS, 1992.

[8] In-Jung Park, Chun-Woo Lee, Ho-Sung Chang, "A Fast Algorithm for Training Multilayer Perceptron Model of Neural Network", NNASP, pp.311 ~ 317, 17-20,August,1993.

[9] Patrick K.Simpson, "Artificial Neural Systems" PERGAMON PRESS, 1990.

[10] 박인정, 김형배, 이상원, "영상정보가 포함된 음성인식에 관한 연구", 단국대학교 멀티미디어 산업기술연구소 논문집 제 1 집, pp.48 ~ 61, 1998.

[11] 김희겸, 이상원, 이천우, 양준승, 박인정, "영상 변화도를 이용한 음성인식," 1999년도 대한전자공학회 멀티미디어연구회 창립총회 및 학술대회 논문집, 1999.

표 1. 음성 입력에 대한 신경망의 출력값
(최대 크기 : 1.0 , 최소 크기 : 0.0)

숫자음 / dB	영	일	이	삼	사	오	육	칠	팔	구
0	0.0990388	0.985946	0.993644	0.994274	0.991455	0.994542	0.992003	0.993719	0.997804	0.993142
10	0.958503	0.984898	0.993644	0.998369	0.976590	0.983558	0.926884	0.993574	0.989585	0.959456
20	0.887395	0.808299	0.993644	0.999685	0.603473	0.368454	0.892311	0.952107	0.981610	0.615538
30	0.877717	0.764210	0.993644	0.999797	0.338487	0.005892	0.732426	0.999926	0.979736	0.013024
40	0.224966	0.949919	0.993644	0.999301	0.085302	0.000066	0.192649	0.999910	0.973441	0.000529

표 2. 잡음에 왜곡된 음성입력, 2 차원 선형예측치 보완 및
히스토그램 보완 기법을 사용시, 신경망의 출력값 비교

숫자음	영			사			오			육			구		
구분 / dB	음성	히스토그램	2 차원	음성	히스토그램	2 차원	음성	히스토그램	2 차원	음성	히스토그램	2 차원	음성	히스토그램	2 차원
0	0.9903	0.9998	0.9986	0.9915	0.9997	0.9995	0.9945	0.9978	0.9817	0.9920	0.9884	0.9868	0.9931	0.9793	0.9968
10	0.9585	0.0012	0.9962	0.9766	0.9604	0.9996	0.9836	0.0030	0.6118	0.9269	0.9838	0.0384	0.9595	0.9793	0.9973
20	0.8874	0.0003	0.9880	0.6035	0.9534	0.9995	0.3685	0.0009	0.0680	0.8923	0.6528	0.0164	0.6155	0.9790	0.9972
30	0.8777	0.0001	0.7858	0.3385	0.9444	0.9993	0.0059	0.0008	0.0461	0.7324	0.0468	0.0031	0.0130	0.9781	0.9969
40	0.2250	0.0000	0.0001	0.0853	0.9316	0.9256	0.0000	0.0007	0.0133	0.0192	0.0211	0.0005	0.0005	0.9745	0.9957