가변 정보율 모델을 이용한 음성인식

김 남 수, 김 경 선, 김 진, 공 병 구, 김 상 룡

삼성종합기술원 음성연구실

Speech Recognition based on Variable Information Rate Model

N. S. Kim, K. S. Kim, J. Kim, B. G. Kong, and S. R. Kim

Speech Technology Lab. Samsung Advanced Institute of Technology

요약

기존의 음성인식에서는, 음성의 모든 구간의 정보적 중요도를 같게 두는 고정 정보율 처리가 일반적이다. 고정 정보율 처리는, 변화가 작은 정 구간을 변화가 큰 단 구간보다 중시하는 경향이 있기 때문에, 음성인식에는 부적절한 요소를 내포하고 있다.

본 논문에서는, 가변 정보율 모델을 제시하여, 음성인식 시, 가변 정보율 처리를 수용하게 하였다. 음성의 각 구간마다 정보율 파라메타를 두어, 확률값 계산에 그 구간의 중요도를 반영하였다. 또한, maximum mutual information (MMI)을 이용하여 정보율 파라메타를 학습시키는 방법을 제안하였다. 화자독립 연속어 인식 실험을 통하여, 가변 정보율 모델을 이용한 방법이 기존의 고정 정보율 방법보다 우수한 인식 성능을 보임을 확인할 수있었다.

1 서 론

통상의 음성인식 기법에서는 입력된 음성신호를 프레임 단위로 불럭화 하여 특징벡터를 추출하는 것이 일반적이다. 추출된 특징벡터들은 인식 방법에 따라 기준패턴 혹은 기준모델과의 유사도 계산에 이용되며, 일정한 시간간격마다 얻어진다. 각각의 특징벡터들은 유사도 측정시, 같은 중요도로 취급되며, 이때문에 유사도의 계산이, 길이가 긴 stationary 부분에 좌우되는 경우가 많다. 그러나, 실제로 인식의 단서가 되는 부분은 길이가 짧은 transient 부분일 경우가 많다는 점에 기초하면, 이는 바람직 하지 못한 현상이라 할 수있다. 따라서, 각 음성구간이 인식의 단서를 제공하는 정도에 따라 그 구간의 중요도를 달리 가져가는 인식 방법이 있어야 한다.

이를위한 몇몇 노력이 있었는데, [1] 에서는 연속되는 유사 특징벡터 열이 있을 경우 하나의 특징벡터만을 취하도록 하였다. 한편, [2] 에서는 음성신호의 energy가 기준 energy에 비하여 특정 한도를 넘는 시간에서만 특징벡터를 추출하였다.

본 논문에서는, 각 음성구간에서의 특징벡터 발생률을 달리하는 가변 정보율 모델을 제시하고 이에의한 음성인식의 방법을 제안한다. 제안된 방법에서는, 기존의 hidden Markov model (HMM) 방식의 음성인식을 수정하여 가변 정보율을 수용할 수있도록 하였다. 우선 전체 음성신호를 일정간격으로 나눈 후 각 구간에서의 특징벡터 발생률을 정보율 파라메타로 설정하였다. 정보율 파라메타는 HMM을 통한 확률값 계산 시 특정 구간의 중요도를 반영한다. 각 구간의 정보율 파라메타 추정은 학습을 통하여 얻어지는데, 이때 학습의 criterion으로는 maximum mutual information (MMI) [3] 이 쓰였다.

2 가변 정보율 모델

일반적으로, 음성신호의 특징벡터 열은, 그림 1과 같이 일정 구간의 음성데이타에 분석창 (analysis window) 을 씌운 다음, 해당 프레임의 특징벡터를 추출하고, 분석창을 일정 길이만큼 이동해 나감으로 얻어진다. 이때, 분석창이 이동하는 길이가 매 프레임마다 일정하면 고정 정보율 분석이라 하고 그렇지 않으면 가변 정보율 분석이라 정의할 수있다.

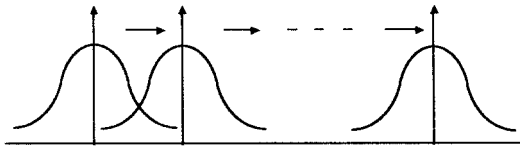

그림 1 분석창 이동에 의한 특징벡터의 추출

고정 정보율과 가변 정보율 분석을 좀 더 자세히 들여다 보기로 한다. 그림 2에서, 각각의 시간 $T_f, 2T_f, \ldots, TT_f$는 특징벡터 추출의 기본 구간을 결정하는데, 각 구간의 길이는 T_f로 일정하다.

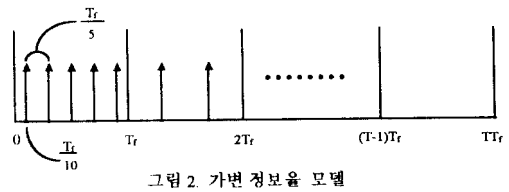

그림 2. 가변 정보율 모델

각 구간에 표시된 화살표와 시간축상 위치는, 그 기본 구간에서 특징벡터 추출 시, 분석창의 중앙이 자리잡는 시간을 나타낸다. 각각의 기본 구간 내에서는, 그 구간에서 추출되어야 하는 특징벡터의 수가 정해지면, 분석창이 자리잡는 위치가 일의적으로 결정된다. 즉, 구간 $[mT_f, (m+1)T_f]$에서 추출되어질 특징벡터의 수를 n이라 하면, 첫번째 분석창의 중앙은 $(m+1/2n)T_f$에 위치하고 T_f/n의 길이만큼 이동하게 된다. 정보율이란, 단위 기본 구간에서 추출되어지는 특징벡터의 수를 나타내는 것으로, 고정 정보율 분석이란 모든 기본 구간에서의 정보율이 서로 같을 때를 뜻하고, 가변 정보율 분석이란 그렇지 않을 때를 말한다.

가변 정보율 모델은 가변 정보율 분석에 근거하는 것으로, 각 기본 구간마다 서로 다른 정보율을 적용한다. 구간 $[(m-1)T_f, mT_f]$에서 추출된 특징벡터 열을 $x_{m1}, x_{m2}, \ldots, x_{mn_m}$이라 하자. 이 때, n_m은 이 구간에서 추출된 특징벡터의 수를 표시한다. 또한, 분석창의 중앙이 $(m-\frac{1}{2})T_f$에 위치했을 때의 특징벡터를 o_m이라 하고, o_m을 이 구간에서의 대표특징벡터라 하자. 대표특징벡터 열 $o_1, o_2, \ldots, o_T$는 통상적인 고정 정보율 분석을 통하여 얻어지는데, 이를 이용하여 가변 정보율 분석으로 얻어지는 특징벡터 열을 근사화 할 수있다. 근사화 방법은 다음과 같다.

1) 0차 근사화

$$x_{mi} = o_m \tag{1}$$

2) 1차 근사화

$$x_{mi} = a\,\frac{2i-1}{2n_m} + b \tag{2}$$

$$a = \frac{1}{2}(o_{m+1} - o_{m-1}) \tag{3}$$

$$b = \frac{1}{2}(o_m + o_{m+1}) \tag{4}$$

3) 2차 근사화

$$x_{mi} = a\,(\frac{2i-1}{2n_m})^2 + b\,(\frac{2i-1}{2n_m}) + c \tag{5}$$

$$a = o_{m+1} + o_{m-1} - 2o_m \tag{6}$$

$$b = \frac{1}{2}(4o_m - 3o_{m-1} - o_{m+1}) \tag{7}$$

$$c = \frac{1}{2}(o_m + o_{m-1}) \tag{8}$$

대표특징벡터 열 $O = o_1, o_2, \ldots, o_T$가 주어졌을 때, 가변 정보율 모델을 이용한 음성인식 방법은 다음과 같다. 우선 λ를 일반적인 HMM이라 하자. 그러면, HMM λ를 통해 대표특징벡터 열 O가 발생하는 확률, $Pr(\Psi_f = O|\lambda)$는 다음과 같이 구해진다.

$$Pr(\Psi_f = O|\lambda) = \sum_{\mathbf{s}} Pr(\Psi_f = O|\mathbf{s}, \lambda)\ Pr(\mathbf{s}|\lambda) \tag{9}$$

$$= \sum_{\mathbf{s}} Pr(\mathbf{s}|\lambda) \prod_{t=1}^{T} Pr(\Psi_{f,t} = o_t|s_t, \lambda) \tag{10}$$

$$= \sum_{\mathbf{s}} Pr(s|\lambda) \prod_{t=1}^{T} \left[\prod_{n=1}^{n(t)} f(x_{tn}|s_t, \lambda)\right]. \tag{11}$$

이 때, $\mathbf{s} = s_1, s_2, \ldots, s_T$는 state 열을 나타내고 $\Psi_f = \Psi_{f,1}, \Psi_{f,2}, \ldots, \Psi_{f,T}$는 대표특징벡터 열을 나타내는 random 변수이다. (11)은 가변 정보율 모델의 수용을 나타내는데, x_{tn}은 t번째 기본구간에서 n번째로 추출되는 특징벡터를 나타내며, $n(t)$는 이 구간에서의 정보율을 말한다. $f(\cdot|s_t, \lambda)$는 state s_t에서의 출력 확률분포를 나타내고 $Pr(\mathbf{s}|\lambda)$는 state 열 s의 확률값을 뜻한다.

(11)에서, 각 기본구간마다 서로 다른 숫자의 특징벡터를 사용하여 확률값을 계산하는 것이, 가변 정보율 모델이 고정 정보율 모델과 다른 점이라 할 수있다. 지금까지 각 기본구간에서의 정보율을 정수로 한정했는데, 이를 0이 아닌 임의의 실수까지로 확장할 수있다. 우선, t번째 기본구간에서의 정보율 $n(t)$가 정수가 아니라고 가정한다. 그러면, 이 구간에서의 가변 정보율 수용방법은 다음과 같다. 먼저, 정보율이 $[n(t)]+1$이라 가정하고 $x_{t1}, x_{t2}, \ldots, x_{t([n(t)]+1)}$을 추출한다. 다음으로, (11)의 확률값 계산 시,

$$Pr(\Psi_{f,t} = o_t|s_t, \lambda) = \prod_{n=1}^{[n(t)]+1} [f(x_{tn}|s_t, \lambda)]^{\frac{n(t)}{[n(t)]+1}} \tag{12}$$

을 적용한다.

3 정보율 파라메타의 추정

가변 정보율 모델을 음성인식에 적용하기 위해서는, 각 구간에서 정보율을 결정할 수있는 조건들을 분류해 내야한다. 여러 종류의 조건들을 선택할 수 있지만, 여기서는 다음의 두가지를 들기로 한다. 첫번째는 state에 의존적인 정보율 이다. 이는 각 state의 정보적 중요도가 서로 다르다는 가정에서 출발하고, (11)의 계산 시, 각 state에서 독자적으로 특징벡터의 추출이 이루어져야 한다. 그렇지만, 만약 (1)의 0차 근사화를 적용한다면, (11)은 단순히 출력 확률값에 가중치를 두는 형태가 된다. state에 의존적인 정보율은 state의 정보적 중요도를 반영하여 각 state의 변별력 향상에 기여하게 된다.

두번째는 대표특징벡터 열에 의존적인 정보율이다. 이는 t번째 기본구간에서의 정보율 결정에 대표특징벡터 열 $\hat{O}_t = o_{t-p}, o_{t-p+1}, \ldots, o_t, o_{t+1}, \ldots, o_{t+q}$을 이용함을 말한다. 이 경우에는 HMM과는 관계없이, 각 기본구간에서의 정보율을 독립적으로 결정한다. 현재 프레임의 대표특징벡터를 포함하여 앞, 뒤 각각 p, q개의 대표특징벡터를 동시에 고려함으로 특징벡터의 변화 궤적을 반영한다. 대표특징벡터 열에 의존적인 정보율의 결정을 위해서는 각 구간 t에서의 $\hat{O}_t$를 벡터 양자화 (VQ) 등을 거쳐 분류해야 하는 과정이 필요하다.

위의, 정보율 결정 조건의 분류에 따라 M 개의 결정 조건 class가 구성되고, k번째 class의 정보율을 α_k라 하자. 정보율 파라메타의 추정을 위해서는 각 구간의 정보적 중요도를 평가하는 척도가 필요한데, 본 논문에서는 음성인식에서 널리 쓰이는 상호정보 (mutual information)를 사용하기로 한다. 따라서, 정보율 파라메타의 추정은 MMI criterion을 바탕으로 이루어 진다. $\Theta = \{n_1, n_2, \ldots, n_M\}$을 전체 정보율 파라메타의 집합이라고 하면, MMI criterion에 의한 파라메타 추정은 다음과 같다.

$$\hat{\Theta}_{MMI} = argmax_{\Theta}\ I(\Theta) \tag{13}$$

$$= argmax_{\Theta} \left\{ \sum_{O_w \in U} \log \left[\frac{P_{\Theta}(\Psi_f = O_w|\lambda_w)}{P_{\Theta}(\Psi_f = O_w)} \right] \right\} . \quad (14)$$

여기서 U는 전체 학습데이타의 집합을 나타내고, λ_w는 실제 O_w를 발생시키는 단어의 HMM을 뜻한다. (13)에서의 $P_{\Theta}(\Psi_f = O_w)$는 O_w의 선 확률을 나타내는데, 이는 N-best 알고리즘을 이용하여 다음과 같이 근사화 할 수있다.

$$P_{\Theta}(\Psi_f = O_w) \approx \frac{1}{N} \sum_{i=1}^{N} P_{\Theta}(\Psi_f = O_w|\lambda_{(i)}) , \quad (15)$$

이 때 $\lambda_{(i)}$는 O_w를 발생시킬 확률이 i번째로 높은 단어의 HMM을 나타낸다.

MMI criterion에 의한 정보를 파라메타 추정을 위해서는, $I(\Theta)$의 정보를 파라메타에 대한 미분치의 계산이 필요하다. 그런데, $I(\Theta)$는 정보율이 정수인 점에서 불연속이 되어, 미분치가 존재하지 않게 된다. 따라서, 정보율이 정수인 부분에서는 미분치를 근사화 하여 결정해야 한다. 이를 바탕으로 정보를 파라메타에 대한 $I(\Theta)$의 미분치를 구하면 다음과 같다.

1) $\lfloor n_k \rfloor < n_k < \lfloor n_k \rfloor + 1$인 경우

$$\frac{\partial}{\partial n_k} P_{\Theta}(\Psi_{f,t}|s_t, \lambda) = \frac{\partial}{\partial n_k} \left\{ \prod_{n=1}^{\lfloor n(t) \rfloor + 1} [f(x_{tn}|s_t, \lambda)]^{\frac{n(t)}{\lfloor n(t) \rfloor + 1}} \right\} \quad (16)$$

$$= P_{\Theta}(\Psi_{f,t} = O_t|s_t, \lambda) \frac{\delta(n(t), n_k)}{\lfloor n(t) \rfloor + 1} \sum_{n=1}^{\lfloor n(t) \rfloor + 1} \log [f(x_{tn}|s_t, \lambda)] \quad (17)$$

이 때 $n(t)$의 결정 조건의 분류가 n_k와 같으면 $\delta(n(t), n_k) = 1$이고 그렇지 않다면 $\delta(n_t, n_k) = 0$이 된다.

2) $n_k = \lfloor n_k \rfloor$인 경우

$$\frac{\partial}{\partial n_k} P_{\Theta}(\Psi_{f,t}|s_t, \lambda) \approx$$
$$P_{\Theta}(\Psi_{f,t} = O_t|s_t, \lambda) \frac{\delta(n(t), n_k)}{n(t)} \sum_{n=1}^{n(t)} \log [f(x_{tn}|s_t, \lambda)] \quad (18)$$

(17)과 (18)을 이용하여

$$\frac{\partial}{\partial n_k} I(\Theta) = \sum_{O_w \in U} \left\{ \frac{1}{P_{\Theta}(\Psi_f = O_w|\lambda_w)} \frac{\partial}{\partial n_k} P_{\Theta}(\Psi_f = O_w|\lambda_w) \right.$$
$$\left. - \sum_{i=1}^{N} \left[\frac{1}{\sum_{j=1}^{N} P_{\Theta}(\Psi_f = O_w|\lambda_{(j)})} \frac{\partial}{\partial n_k} P_{\Theta}(\Psi_f = O_w|\lambda_{(i)}) \right] \right\} \quad (19)$$

를 구할 수있다.

$I(\Theta)$와 미분치가 구해졌을 때, 가장 생각하기 쉬운 파라메타 추정 방법은 steepest ascent 기법으로 다음과 같이 수행된다.

$$n_j|_{\tau+1} = n_j|_{\tau} + \Delta \left. \frac{\partial I(\Theta)}{\partial n_j} \right|_{\tau} , \quad (20)$$

이 때 $n_j|_{\tau}$는 τ번째 iteration에서의 n_j 값이고, Δ는 stepsize를 나타낸다. Steepest ascent 방법은 간편히 사용할 수있지만, 일반적으로 수렴속도가 느리고 stepsize의 결정이 어렵다. 이에대한 대안으로 extended Baum 알고리즘 [4] 을 이용한 파라메타 추정 방법을 제안한다. 먼저, 각각의 정보율 파라메타가 다음의 제약식을 만족하도록 한다.

$$\sum_{j=1}^{M} n_j = MR , \quad (21)$$

이 때, R은 평균 정보율을 나타내는 상수이다. 음성인식에 있어서는, 정보율의 절대적 크기보다는 상대적 비교가 중요함으로 (21)은 의미를 지닌다. (21)의 제약식을 바탕으로, extended Baum 알고리즘을 사용한 파라메타 추정 방법은 다음과 같다.

$$n_j|_{\tau+1} = MR \frac{n_j|_{\tau} \cdot D_j|_{\tau}}{\sum_{j=1}^{M} n_j|_{\tau} \cdot D_j|_{\tau}} , \quad (22)$$

여기서 $D_j|_{\tau}$는 각각의 정보율에 대한 $I(\Theta)$의 미분치에 의하여 결정된다. (22)에서 $D_j|_{\tau}$는 $\frac{\partial I(\Theta)}{\partial n_j}|_{\tau}$의 값을, 모든 j에 대하여 음수가 되지 않게하고, 가장 작은 미분치와 가장 큰 미분치와의 비를 일정 한도 이하가 되도록 보장한 값이다. 이에따라, 허용한도를 $\Im$이라 했을 때, $D_j|_{\tau}$는 다음과 같이 구해진다.

1) 모든 j에 대하여 $\frac{\partial I(\Theta)}{\partial n_j}|_{\tau} > 0$이고

$$MAX/MIN < \Im ,$$
$$MAX = \max_{j=1}^{M} \left\{ \left. \frac{\partial I(\Theta)}{\partial n_j} \right|_{\tau} \right\}$$
$$MIN = \min_{j=1}^{M} \left\{ \left. \frac{\partial I(\Theta)}{\partial n_j} \right|_{\tau} \right\}$$

일 때,

$$D_j|_{\tau} = \left. \frac{\partial I(\Theta)}{\partial n_j} \right|_{\tau} . \quad (23)$$

2) 그 외의 경우

$$D_j|_{\tau} = \left. \frac{\partial I(\Theta)}{\partial n_j} \right|_{\tau} + \varepsilon , \quad (24)$$

여기서

$$\varepsilon = \frac{MAX - \Im MIN}{\Im - 1} . \quad (25)$$

4 화자독립 연속어 인식 실험

인식 실험에 사용된 어휘는 달, 요일, 날짜와 시간을 표시하는 102개의 단어로 구성되었다. 90명 (남자 43명 여자 47명)의 화자가 20-30 문장을 발음하여 학습과 인식에 쓰일 database를 구축하였다. 70명 (남자 33명 여자 37명)의 음성데이타가 학습에 쓰였는데, 학습에 사용된 총 단어수는 5122개 였다. 나머지 20명의 데이타는 인식에 쓰였으며 이때의 총 단어수는 1448개 였다. 각각의 음성신호는 4.5 kHz의 대역폭을 갖는 low-pass filter를 통과하여 16 kHz의 rate로 sampling 되었다. 특징벡터로는 12차의 LPC cepstrum 계수와 이들의 1차 변화를 나타내는 delta cepstrum 계수가 사용되었으며, Codebook size 256인 두개의 codebook이 각각의 특징벡터에 대하여 구성되었다. 27개의 음소모델이 인식의 기본 단위로 선택되었고, 각각의 음소에 3개의 state로 구성된 이산분포 HMM을 사용하였다. HMM의 파라메타는 maximum likelihood (ML) criterion에 근거하여 학습되었다.

가변 정보율 모델을 사용하지 않았을 때의 단어 인식률은 73.0 %였다. 가변 정보율 모델을 이용한 음성인식 실험에서는 state에 의존적인 정보율이 쓰였는데, 각 state마다 정보율 파라메타를 설정하였다. 인식 시간의 단축을 위하여 0차 근사화를 통하여 단순히 출력 확률값에 가중치를 두는 방법을 선택하였고, 정보율 파라메타의 학습 시 고정된 $\Im$을 사용하였다. $\Im$의 값에 따른 인식률의 변화는 표 1에 나와있다. 인식 결과를 살펴보면 $\Im$이 작을 때는 iteration 간의 인식률 변화가 작고 $\Im$이 클때는 매우 심한 인식률 변화를 보임을 알 수있다. 가장 높은 인식률을 보인것은 $\Im = 1.2$ 일 때로 가변 정보율을 사용하지 않았을 때의 단어 오인식률을 19.0 % 감소시켰다.

표 1: 가변정보율 모델을 이용한 인식 실험 결과.

iteration	인식률 (%)		
	$\Im = 1.2$	$\Im = 1.5$	$\Im = 2.0$
1	75.6	74.6	67.5
2	77.0	75.2	70.0
3	76.7	75.1	71.5
4	78.0	75.0	66.7
5	76.2	73.8	77.1
6	75.6	76.3	69.2

5 결 론

본 논문에서는 가변 정보율 모델을 이용한 음성인식의 방법을 제안하였다. 가변 정보율 모델을 통하여 각 음성구간의 중요도가 인식에 반영될 수있었다. state에 의존적인 정보율 모델을 이용한 음성인식 실험에서 제안된 방법은 기존의 고정 정보율 모델에의한 단어 오인식률을 19.0 % 까지 감소할 수있었다. 대표특징벡터 열에 의존적인 정보율 모델을 적용할 경우 더 높은 인식 성능 향상이 기대된다.

참고 문헌

[1] R. M. Schwartz et al., "Context-dependent modeling for acoustic-phonetic recognition of continuous speech," *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL., pp. 1205-1208, Mar. 1985.

[2] K. M. Ponting and S. M. Peeling, "The use of variable frame rate analysis in speech recognition," *Computer Speech and Language*, Vol. 5, pp. 169-179, 1991.

[3] L. R. Bahl et al., "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. of IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan, pp. 49-52, April 1986.

[4] P. S. Gopalakrishnan et al., "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. Inform. Theory*, Vol. 37, pp. 107-113, Jan. 1991.