인식률 향상을 위한 프레임 가변 이동방식 한국어 음성인식시스템 설계
Korean Speech Recognition System Using Variable Distance Method of Frame Movement for Improving Recognition

한승진(인터넷경영정보과),이재호(인터넷경영정보과)
Sung-Jin Han(Dept. of Internet and Management Information System)
Jae-Ho Lee(Dept. of Internet and Management Information System)

Key Words : 프레임 가변 이동방식

요약 : 음성 인식기는 인간이 발음할 수 있는 모음과 자음을 인식하는 시스템이다. 그러나 모음과 자음의 발화 길이 차이를 고려하지 않고, 평균 음절속도만을 측정하기 때문에 인식률이 저하된다. 인식 오류율이 모음과 모음사이에 있는 자음소 경계구간에서 상대적으로 높으므로 본 논문에서는 발화 길이를 측정하고, 이 측정된 정보를 이용하여 인식 과정 시에 보상하는 방법인 프레임 가변거리 이동방식을 이용한 한국어 음성인식시스템을 제안한다. 훈련된 각 음소 모델을 이용한 연속음 인식 실험결과 가변 거리방식을 적용하였을 때 기존 방식에 비해 5%의 인식률 향상을 보였다.

1. 서론

음성의 발화속도는 연속음 인식에 커다란 영향을 미치는 것으로 알려져 왔다[1]. 또한 대부분의 음성분석은 화자가 주의 깊게 발음한 데이터에 대한 것이다. 따라서 자연 발화시의 음성은 이러한 음성과 다소 차이를 보인다. 이 때문에 기존의 연속음 인식 시스템에서는 확률모델을 이용하여 전체속도에 대한 평균에 의해 발화속도를 고려하거나[2] 언절에 따른 속도를 분석하여 처리를 하지만[3], 음절의 길이와 모음과 자음의 발화 길이 차이를 고려하지 않았고, 이로 인한 인식율의 저하를 피할 수 없었다[4]. 따라서 본 논문에서는 단모음구간과 장모음구간, 그리고 자음구간에서 각기 다른 프레임 이동거리를 적용하여 오류율을 줄이는 시스템을 설계하고 구현한다.

2. 프레임 고정거리 이동방식

음성은 시간에 따라 변화하는 특징을 가지고 있으나, 20~30ms정도의 짧은 시간간격 동안은 시 불변이라고 가정할 수 있다. 이를 바탕으로 평균 20ms의 프레임단위로 특징벡터를 추출하고 프레임사이의 손실된 정보를 보상하기 위해 10ms단위로 프레임을 이동한다.

3. 발화속도에 따른 프레임 이동방식

언절의 길이와 언절에 포함된 음절의 개수로 발화속도를 계산하여 프레임 이동거리를 정한다.

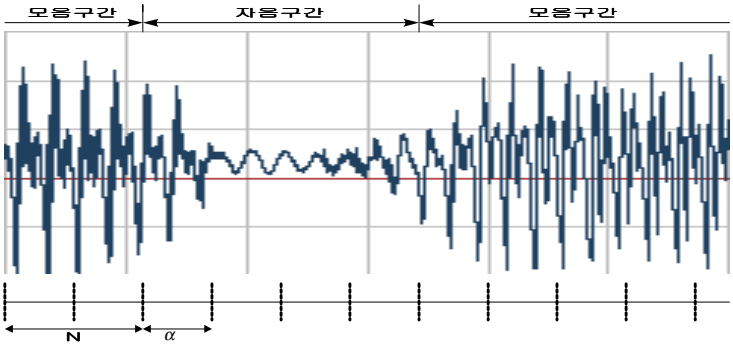

그림 1. 프레임 고정거리 이동방식

$$\text{평균음절속도} = \frac{\text{음절의 개수}}{\text{발화단위의 총 크기}} \cdot f_s$$

f_s : 표본화 주파수

그러나 한 언절에는 동일한 프레임 이동길이를 적용하므로 긴 음절과 짧은 음절이 같이 포함된 경우에는 발화속도를 정확히 측정할 수 없어 효과적인 프레임 이동거리를 정할 수 없는 문제점이 있다.

4. 프레임 가변거리 이동방식

본 논문에서 제안하는 프레임 가변거리 이동방식은 포만트 주파수의 변화율에 의해 모음구간을 검출한 후, 이를 기준으로 발화길이가 길고 주파수 변화가 적은 모음구간은 10ms이상의 간격으로 프레임을 이동시켜 특징벡터를 추출하고, 상대적으로 발화길이가 짧고 주파수 변화가 큰 자음구간에서는 10ms보다 적은 간격으로 프레임을 이동시켜 특징벡터를 추출한다. 즉, 자음구간에서는 모음구간보다 프레임 이동거리를 좁혀 특징벡터를 보다 많이 추출한다.

4.1 모음구간 검출

음성의 기본 파라메터 중 포만트 주파수는 모음영역에서 에너지가 집중적으로 나타나는 특성을 가지고 있다. 포만트 주파수는 LPC(Linear Predictive Coefficient)를 이용하여 구할 수 있다. LPC계수를 a_i라 하면 이때 $A(z)$의 역필터는 식(1)과 같다.

$$A(z) = A_s(z) \cdot A_u(z)$$
$$= \prod_{i=1}^{q/2}(1 - b_i z^{-1})(1 - b_i^{*z^{-1}}) \prod_{i=q+1}^{p}(i - c_i z^{-1}) \quad (1)$$

LPC 분석에 의해 $S(z)$의 다항식은 $B(z)$가 되고, 최대 q/2개의 포만트를 가진다. LPC분석에서는 불안정근의 소거에 의한 차수의 감소로 충분히 높은 차수로 분석할 때 충실도를 갖는 분석이 가능하다[5].
포만트는 성도의 공진 특성을 모델링 한 것으로, 공진 주파수의 크기와 대역폭의 추정이 필요하다. 포만트 주파수 F는 식(1)의 안정화된 근 $B(z)$에서 실수부와 허수부의 분리에 의해 다음과 같이 구해진다.

$$F_i = \frac{1}{2\pi T}\tan^{-1}\left[\frac{Im(z_i)}{Re(z_i)}\right] \qquad (2)$$

$$B_i = -\frac{1}{2\pi T}\log\left[Re(z_i)^2 + Im(z_i)^2\right] \qquad (3)$$

여기서 T는 표본화 시간이다. 구해진 포만트 주파수와 대여폭에서 크기순으로 5개를 선택하여 포만트 주파수로 설정한다.
식(4)에 의해 포만트 주파수 변화율을 계산하여 네 개의 포만트가 모두 변하는 구간을 판단하여 모음구간을 측정한다.

$$M(n) = M(n-1) + \sum_{i=1}^{fn}\{F_i(n+1) - F_i(n)\} \quad (4)$$

변화율이 임계치를 넘는 구간에서 모음이 발생되었다고 판단할 수 있다.

4.2 음소의 길이

음절을 구성하는 음소는 음절수, 음절의 위치 등에 따라 지속시간의 변화를 가진다[6]. 음절수가 증가함에 따라 음소별 지속시간은 감소하는 경향을 보이는데, 단음절에서 3음절까지의 음소별 지속시간은 현저한 차이를 보이고, 그 이상의 경우 지속시간 감소율은 줄어든다. 그리고 첫 음절과 중간음절의 음소지속시간보다 마지막 음절에서 자음은 20% 더 길어지고, 모음은 80% 길어진다[7]. 음성인식에서 사용되는 음성은 3음절이상이고 오류발생은 음절의 중간에서 많이 발생하므로, 자음과 모음의 평균 지속시간비율을 구하기 위해서는 마지막 음절을 제외하고, 나머지 음절에서 자음, 모음의 지속시간을 구한다.
임의의 화자 10명의 자연 발화를 통계적으로 분석해 본 결과, 자음 지속시간은 85.9ms이고 모음은 140ms 이었다. 분석값을 토대로 자음과 모음의 길이 비율은 약 4 : 6을 보임으로써 모음구간이 자음구간보다 길다는 것을 알 수 있다.

4.3 프레임 이동거리 계산

모든 음소에서 동일한 비율의 특징벡터 추출을 위하여 모음과 자음의 평균 길이 비율을 사용하였다.

자음 : 모음 $= \alpha : \gamma$

$\alpha + \gamma =$ 음소이동거리 · 2

$\beta = (\alpha + \gamma)/2$

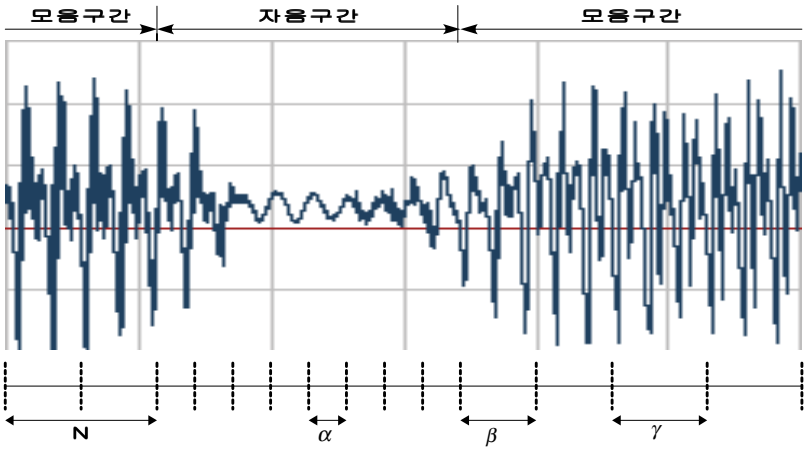

그림 2. 프레임 가변거리 이동방식

통계에 의한 자음과 모음의 길이 비율과 기본 이동 거리 10ms의해 구간에 따라 α, β, γ를 이동 비율로 정한다.

구간 $S(n)$와 $S(n+1)$사이에서 프레임의 이동거리는 다음 식(5)에 의해 계산된다.

$$S(n+1)=S(n)+\alpha,\ \left\{ \begin{array}{l} S(n) \not\ni Vm_i \\ S(n) \ni Cm_i \end{array} \right\} \qquad (5)$$

$$S(n+1)=S(n)+\beta,\ \{S(n) \ni (Vm_i \wedge Cm_i)\}\ S(n+1)=S(n)+\gamma,\ \left\{ \begin{array}{l} S(n) \ni Vm_i \\ S(n) \not\ni Cm_i \end{array} \right\}$$

Vm_i : 모음 구간

Cm_i : 자음 구간

$S(n)$: n 번째 프레임 구간

프레임 구간 $S(n)$내에 모음구간이 없을 경우와 프레임 구간 $S(n)$이 모음구간과 자음구간을 모두 포함할 경우, 그리고 프레임 구간 $S(n)$내에 자음구간이 없을 경우를 각각 구분하여 프레임 이동간격을 조절한다

4.4 프레임 가변거리 이동방식 음성인식 시스템

입력 음성에 대하여 에너지와 영교차율로 음성을 검출한다. 검출된 음성의 포만트 주파수를 구한 후, 포만트 주파수 변화율에 의해 음절의 모음구간을 결정한다. 모음구간을 중심으로 식(5)에 의해 각 구간의 프레임 이동율이 계산된다. 음성 구간의 프레임 이동율이 정해지면 음성의

전역에 대한 음성 손실을 보상하고자 전역 필터로 다음과 같이 표현되는 선 강조 함수를 적용한다.

$$H(z) = 1 - 0.95z^{-1} \qquad (6)$$

추출된 프레임의 양끝의 잡음을 제거하기 위해 해밍창을 통과시키고. 인간의 청각적 특성을 고려한 12 차 MFCC(Mel Frequency Cepstrum Coefficient)를 추출하여, 이 특징벡터를 HMM의 입력벡터로 사용한다. 제안한 시스템의 전체 흐름도를 그림 3에 나타내었다.

5. 실험 및 결과

제안한 시스템의 성능을 평가하기 위한 음성DB는 20대 남녀 각 10명의 화자가 47개의 음소를 포함한 100개의 단어를 5번씩 발음한 데이터로 구축하였고, 이 DB로부터 HMM을 학습시켰다. 실험은 잡음이 고려되지 않은 실험실에서 음성DB구축에 참여하지 않은 20명의 남녀가 독립단어와 연속음을 두 번씩 발성하였다. 발화속도는 개인에 따라 다르나 의식적으로 빠르게 하거나 느리게 하지는 않았다. 인식율과 연산량을 비교하기 위하여 10ms간격의 프레임 고정거리이동 방식과 제안한 프레임 이동의 가변거리방식을 비교하였다.
실험 결과 프레임 고정거리방식보다 독립단어 2%, 연속음 5%의 인식율이 향상되었다. 연산량의 경우 독립단어 6.75%, 연속음 5.8%의 증가율을 보였다.

알고리즘 1. 이동거리 계산 알고리즘

```
if (s[n]  ∋Vm)
        if (s[n] ∋  Cm)   이동거리= β
        else  이동거리= γ
else
        이동거리= α
s[n+1]=s[n]+이동거리
```

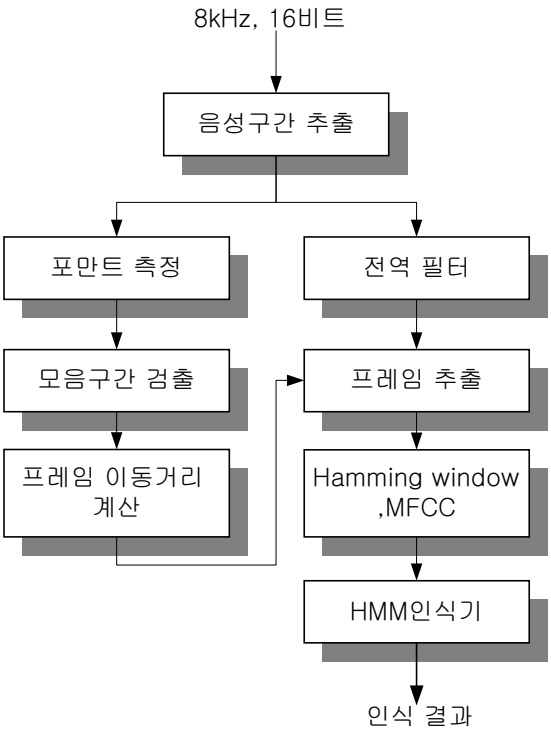

그림 3. 프레임 가변거리 이동방식 음성인식 시스템

표 1. 인식결과

	인식율		연산량	
	독립단어	연속음	독립단어	연속음
고정거리 이동방식	84%	51%	74.4프레임	639.4프레임
가변거리 이동방식	86%	56%	79.5프레임	650프레임
향상율	2%	5%	6.75%	5.8%

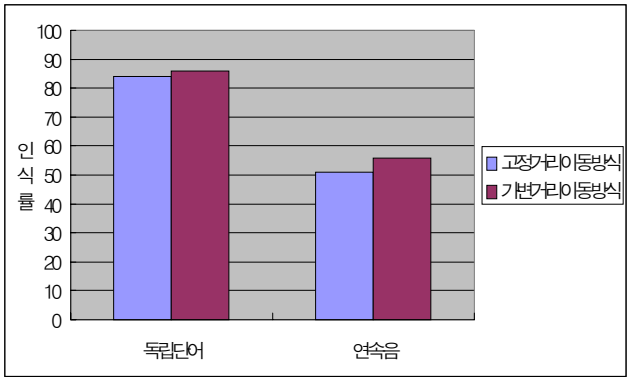

그림 4. 인식결과 비교 그래프

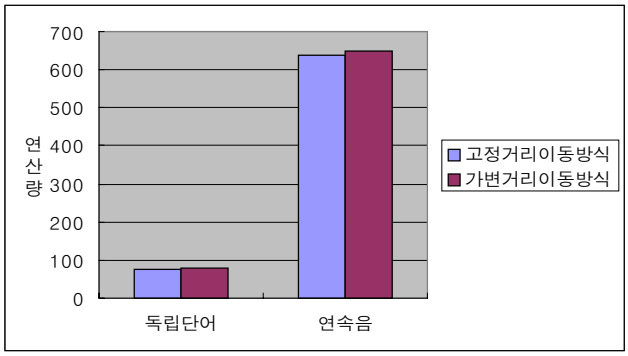

그림 5. 연산량 비교 그래프

6. 결론

본 논문에서는 인식율을 향상시키기 위하여 새로운 프레임 이동 방식을 제안하였다. 기존의 프레임 이동방식은 모든 음성에 대하여 동일하게 프레임을 이동시켜 특징벡터를 추출하지만 이는 자음과 모음의 길이를 고려하지 않아 다양한 음소길이가 포함된 음성에서는 정확한 특징벡터를 추출할 수 없다. 제안한 프레임 가변거리 이동방식은 자음과 모음에 따라 특징벡터의 추출 수를 다르게 하여 다양한 음소길이를 포함하고 있는 음성에서의 인식율을 높일 수 있었다.
독립단어와 연속음으로 발화실험을 수행한 결과 연속음에 대해서는 5%의 인식율 향상을 보였는데, 이와 같은 결과는 제안한 시스템이 독립단어에 적용할 때 보다 다양한 음소길이가 포함되어 있는 연속음에 적용할 때 더욱 적합하다는 것을 알 수 있다.
반면에 연산량은 5.8%증가하는 결과를 보여 향후 연구 과제로는 증가된 연산량을 줄이는 연구와 발화속도에 따른 음소변화율을 이용한 인식율 향상 연구가 필요하다.

참고문헌

(1) David Burshtein, "Robust Parametric Modeling of Durations in Hidden Markov Models," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Vol. 1, pp. 683-686, 1995.
(2) Lawrence Rabiner, Biing-Hwang Jung, *Fundermantals of Speech Recognition*, 1993.
(3) 김재범, 박찬규, 한미성, 이정현, "발화속도 적응적인 한국어 연속음 인식기," 한국어 정보처리학회논문지, 제4권 제6호, 1997. 6월, pp.1531-1540.
(4) D. O'Shaughnessy, "Timing Patterns in Fluent and Disfluent Spontaneous Speech," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Vol. 1, pp. 600-603, 1995.
(5) J.d.Markel,A.H.Gray, *Linear Prediction of Speech*, Springer Verlag, New York,1976.
(6) John Clark, Colin Yallop, *An Introduction To Phonetics & Phonology*, Blackwell, 1995.
(7) 최영익, 권철홍, "자연스러운 여성 합성음을 위한 지속시간 규칙에 관한 연구," 한국음향학회 학술발표대회 논문집, 제18권 1호, 1999년, pp.3-6.