2001年度 第14回 信號處理合同學術大會 論文集 第14卷 1號

Lip-synch application을 위한 한국어 단어의 음소분할

강 용 성, 고 한 석
고려대학교 전자공학과
서울시 성북구 안암동 5가 1번지

The segmentation of Korean word for the lip-synch application

Yongsung Kang, Hanseok Ko
Department of Electronics Engineering, Korea University
5ka-1 Anam-Dong Sungbuk-Ku, Seoul 136-701, Korea
Tel: +82-2-926-2909, Fax: +82-2-3291-2450
yskang@ispl.korea.ac.kr, hsko@korea.ac.kr

Abstract

본 논문은 한국어 음성에 대한 한국어 단어의 음소단위 분할을 목적으로 하였다. 대상 단어는 원광대학교 phonetic balanced 452단어 데이터 베이스를 사용하였고, 분할 단위는 음성 전문가에 의해 구성된 44개의 음소셋을 사용하였다. 음소를 분할하기 위해 음성을 각각 프레임으로 나눈 후 각 프레임간의 스펙트럼 성분의 유사도를 측정한 후 측정한 유사도를 기준으로 음소의 분할점을 찾았다. 두 프레임 간의 유사도를 결정하기 위해 두 벡터 상호간의 유사성을 결정하는 방법중의 하나인 Lukasiewicz implication을 사용하였다. 본 실험에서는 기존의 프레임간 스펙트럼 성분의 유사도 측정 을 이용한 하나의 어절의 유/무성음 분할 방법을 본 실험의 목적인 한국어 단어의 음소 분할 실험에 맞도록 수정하였다. 성능평가를 위해, 음성전문가에 의해 손으로 분할된 데이터와 본 실험을 통해 얻은 데이터와의 비교를 하여 평가를 하였다. 실험결과 전문가가 직접 손으로 분할한 데이터와 비교하여 32ms이내로 분할된 비율이 최고 84.76%를 나타내었다.

1. 서론

음성분할은 화자인증 및 음성인식 시스템의 성능 향상이나, 대용량 음성 데이터베이스 구축과 같은 용도로 사용되어 질 수 있다. 본 논문에서는 립싱크 시스템의 전처리로서 연속된 음성입력에서 음소를 구분하기 위한 방법으로 음성분할 실험을 하였다.

음성을 분할하는 방법은 크게 두 가지로 나눌 수 있다. 첫번째는 분할할 음성에 대한 정보를 미리 알고 있어서 그 정보를 이용하여 분할 하는 방법이고, 두번째는 분할할 음성에 대한 정보를 모르는 상태에서 음성이 포함하고 있는 정보만을 이용하여 음성을 분할 하는 방법이다. 첫번째 방법으로는 음성이 어떤 음성을 발음한 것인지 알고 그것을 이용하는 방법, 음소의 정보를 미리 알고 있어서 그 정보를 이용하여 분할을 하는 방법 등이 있다. 두번째 방법으로는 음성의 모델을 세운 후 시간의 변화에 따라 모델의 변화를 측정하는 GLR(General likelihood ratio)방법[3], 음성의 스펙트럼정보의 시간에 따른 변화를 측정하는 방법[1] 등이 있다. 본 실험에서 사용하는 방법은 시간에 따른 음성의 스펙트럼 정보의 변화를 측정해서 음성을 분할하는 방법을 사용하였다.

기존 유사도 측정을 이용한 음성분할 방법은 음성신호를 각 프레임별로 나눈 후 인접한 프레임 간의 스펙트럼 정보의 유사도를 Lukasiewicz 와 Gains implication을 이용하여 결정한 후 그 유사도의 변화를 기준으로 음성의 유성음과 무성음을 구분하는 방법을 사용하였다.

본 실험에서는 기존의 프레임간 스펙트럼 성분의 유사도를 바탕으로 한 유/무성음 분할 실험[1]을 바탕으로 하였다. 하지만, 기존 실험의 음성분할은 중국 만다린어 음절내의 유/무성음 분할을 목적으로 하고 있기 때문에 본 실험의 목표인 단어내의 음소분할에 바로 적용할 수 없다. 따라서, 본 실험에 맞도록 기존 방법을 수정하였다. 첫째, 프레임 간의 유사도 측정 방법을 고주파 영역과 저주파 영역을 구분하지 않고 하나의 유사도 측정방법, Lukasiewicz implication을 사용하였다. 이때, 이 방법이 고주파 영역에서 유사도를 충분히 반영하지 못하는 단점을 보완하기 위하여 음성을 pre-emphasis 처리 하였다. 둘째, 실험의 목적에 맞도록 decision rule의 보완이 이루어졌다. 음절을 대상으로 할 때와는 달리 단어를 대상으로 음소분할을 할 때에는 음절과 음절간의 휴지기간에 대한 고려가 있어야 하며, 또한, 한 단어 내에는 두 개 이상의 분할점이 존재하므로 기존의 가장 큰 유사도의 변화점을 찾는 방식으로는 단어 내에서의 음소분할을 수행할 수 없다. 따라서, 본

논문에서는 단어의 음소분할에 적합한 decision rule을 제시한다.

본 논문의 구성은 2장에서 기존에 사용된 유사도를 측정하는 방법에 대해 서술하고, 3장에서 본 논문에서 실험목적에 맞도록 제안한 방법을 서술한다. 그리고 4장과 5장에서 실험 및 실험 결과를 분석하고 6장에서 결론을 맺는다.

2. 기본적인 Lukasiewicz implication을 이용한 프레임간 유사도 측정

2.1 Clustering method.

프레임간의 유사도를 측정하기 위해, 각 프레임 마다 데이터 세트를 다음과 같이 구성한다.

$$(x_k, y_k)\ ,\ k = 1,2,\cdots,N\ ,$$
$$\text{where } x_k \in [0,1]^n \text{ and } y_k \in [0,1]^n$$

두개의 데이터 세트 (x_k, y_k), (x_l, y_l)가 같은 cluster에 존재하려면, 다음 두 가지의 성질을 만족해야 한다[2].

(1) 같은 차원의 x_k와 x_l이 유사(similar)하고, 또한 y_k와 y_l이 역시 유사해야 한다.

(2) 특성을 나타내는 인덱스의 지향성 요소는 반드시 함수의 방향을 반영해야 한다. 즉, x_k와 x_l이 매우 유사하고 y_k와 y_l이 매우 다른 경우는 서로 다른 cluster에 속하고, 마찬가지로, x_k와 x_l이 약간 다르고, y_k와 y_l이 매우 유사한 경우는 같은 cluster에 속한다.

유사도(similarity)는 다음과 같이 표현된다. Membership value a와 b의 유사도는,

$$a \equiv b = \frac{1}{2}[\min(a \to b, b \to a) + \min(\bar{a} \to \bar{b}, \bar{b} \to \bar{a})]$$
$$\text{where } a, b \in [0,1],\ \bar{a} = 1 - a,\ \bar{b} = 1 - b$$

와 같다. 여기서, "$\to$"는 multivalued implication을 나타낸다. 만일 membership value a와 b의 값이 같다면, 같은 정도를 나타내는 equality index는 1과 같다. Implication은 매우 다양한 방식으로 정의 된다. 그 중 본 실험에서는 다음 식의 Lukasiewicz implication 방법을 사용한다.

$$a \to b = \begin{cases} 1 & \text{if } a \le b \\ 1 - a + b & \text{otherwise} \end{cases} = \min(1, 1 - a + b)$$

2.2 프레임간의 유사도 측정

음성신호가 입력되면, Hamming 윈도우를 사용하여 신호를 각각 프레임으로 나눈다. 한 프레임 내의 신호를 critical-band mapping하면 데이터열 $\{c(i,k),\ 1 \le i \le N,\ 1 \le k \le C\}$이 얻어진다. 여기서, N은 음성 데이터의 전체 프레임의 수, C는 critical-band의 수를 나타낸다. 인접 프레임 $c(i,k)$와 $c(i+1,k)$의 유사도를 측정하기 위한 clustering-method measurement로 $[c(i,1) \equiv c(i+1,1)]$, $[c(i,2) \equiv c(i+1,2)]$, $\cdots$, $[c(i,C) \equiv c(i+1,C)]$을 각각 구한다. 여기서, "$\equiv$"의 의미는 양쪽 member의 유사도를 나타낸다. $M(i)$는 i번째 프레임과 $i+1$번째 프레임간의 유사도를 나타낸다.

$$M(i) = \frac{1}{C}\sum_{k=1}^{C}[c(i,k) \equiv c(i+1,k)],\ 1 \le i \le N-1$$

각각의 $[c(i) \equiv c(i+1)]$을 구하기 위하여, 위의 유사도 식에 따라 전개하면,

$$c(i) \equiv c(i+1) = \frac{1}{2}\{\min[c(i) \to c(i+1), c(i+1) \to c(i)]$$
$$+ \min[\overline{c(i)} \to \overline{c(i+1)}, \overline{c(i+1)} \to \overline{c(i)}]\}$$

과 같이 나타낼 수 있다. 다음 Lukasiewicz implication 식을 위의 식에 대입하면, 인접 두 프레임간의 유사도 $M(i)$를 구할 수 있다.

Lukasiewicz implication:

$$c(i) \to c(i+1) = \begin{cases} 1, & \text{if } c(i) \le c(i+1) \\ 1 - c(i) + c(i+1), & \text{otherwise} \end{cases}$$
$$= \min[1, 1 - c(i) + c(i+1)]$$

3. 제안한 방법

기존 논문의 방법은 음성신호의 고주파 영역에서 일반적으로 그 에너지가 저주파에 비해 상대적으로 낮기 때문에, 고주파 부분에서 유사도를 측정하는 방법으로 Lukasiewicz implication 대신 비교하는 두 값이 작을 경우 효과적으로 그 유사도를 나타내는데 유용한 Gains implication을 사용하였다. 하지만, 프레임간의 유사도를 측정하는데 있어서, 두 가지의 유사도 측정방법을 적용하는 기준이 확실하지 않고, 고주파 부분에서 신호의 크기가 커진다면 Gains implication방법을 사용하여 고주파 영역의 유사도를 측정할 필요가 없어진다. 따라서, 본 논문에서는 유사도 측정을 하기 전에 음성신호에 대하여 pre-emphasis 처리를 하여 고주파 영역의 신호를 강화한 후, Lukasiewicz implication 한가지만을 사용하여 유사도 측정을 하였다. 다음은 pre-emphasis식을 나타낸다.

$$H(z) = 1 - az^{-1},\quad 0.9 \le a \le 1.0$$
$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1)$$

where, $\tilde{s}(n)$: pre-emphasis signal,
$s(n)$: input signal

기존의 실험은 그 목적이 하나의 중국 만다린어 음절에서의 자음/모음 분할을 목적으로 하였다. 따라서 만다린어 한음절의 특성상 하나의 음절은 자음+모음 혹은 모음으로만 이루어져 있으므로, 구분을 위해서 전체 프레임의 유사도 중 가장 큰 차이를 보이는 프레임을 찾으면 되었다. 하지만, 본 논문의 목적은 한국어 단어에서의 음소단위 분할을 목적으로 하고 있기 때문에, 전체 음성구간에 여러 개의 음소들이 존재하며 음절 사이의 휴지구간 또한 존재하고 있기 때문에, decision rule의 수정이 필요하다. Decision rule의 수정을 위해 음성은 시간에 따라 천천히 변하기 때문에[4],

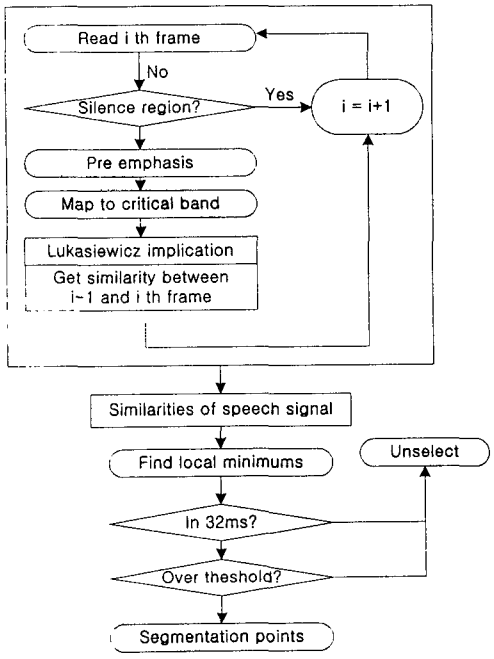

그림 1 전체 시스템 블록도

유사도 역시 갑자기 줄어들지 않고 서서히 줄었다가 다시 서서히 높아진다는 가정을 하였다. 따라서, 위의 가정에 따라 전체 입력된 음성구간 내에 두개 이상의 음소가 존재하기 때문에, 유사도의 local minimum들을 찾으면 음소의 경계를 찾을 수 있다. 하지만 높은 유사도 값을 가지는 local minimum들이 음소가 지속되는 구간에서도 나타난다. 이런 local minimum들을 제거하기 위해, 값을 변화시키면서 반복적인 실험을 바탕으로 threshold를 결정하여 이런 local minimum들을 제거하였다. 그리고, 음소의 지속시간을 고려하여 32ms이하에 중복되어 있는 local minimum들은 그 중 더 작은 값을 선택하도록 하였다.

본 실험에서 사용된 decision rule은 다음과 같다.
(1) 전 프레임에 걸쳐 유사도의 local minimum들을 찾는다.
(2) 음소의 최소 지속시간을 고려하여, 음소 지속시간보다 더 짧은 구간에 속해있는 local minimum을 제거한다. 이때, 제거되는 local minimum은 더 큰 유사도를 나타내는 값으로 한다.
(3) 이렇게 찾은 local minimum들 중 threshold보다 큰 유사도를 가지는 local minimum을 제외시킨다.

또한 단어의 전후 혹은 음절사이의 무음 구간에서는, 스펙트럼 정보가 일정하게 나타나지 않기 때문에 잘못된 분할을 행하게 된다. 이런 오류를 줄이기 위해, 전처리 과정으로 무음구간을 찾아내는 과정이 필요하다. 본 실험에서는 무음 구간을 찾아내기 위해 프레임내의 에너지의 크기를 이용하였다. 한 프레임내의 에너지의 크기가 threshold보다 작으면, 그 프레임은 무음 구간으로 간주하여 음소분할을 수행하지 않았다. 다음 그림 1은 전체 시스템의 블록도이다.

표 1 실험 결과 : 기본 방법과 제안한 방법의 결과와 손으로 분할한 결과와의 비교

	제안한 방법		기존 방법	
Threshold	0.95	0.93	0.8	0.85
In 16ms	2249	2205	1863	2247
In 32ms	712	722	485	316
Lost	536	570	1149	734
Over	992	809	2747	6486
16ms(%)	64.31%	59.02%	53.27%	68.15%
32ms(%)	84.76%	83.70%	67.14%	79.01%

4. 실험 환경

실험에 사용된 음성 데이터는 원광대학교 phonetic balanced 452단어 DB 중 남성화자 6명이 1회 발화한 452단어 데이터를 사용하였다. 각각의 화자가 발음한 음성에는 분할 대상 음소가 균일하게 분포되도록 구성되었다. 음소분할의 단위는 음성전문가에 의해 구성된 44개의 음소를 단위로 사용하였다. 각 프레임의 크기는 32ms, 프레임간 overlap구간은 16ms, Hamming 윈도우를 사용하였다. Pre-emphasis 파라메터는 0.97, Critical band는 18개의 크기를 가지는 Mel bank를 사용하였다. 성능평가를 위해, 음성전문가가 손으로 직접 분할한 데이터와 본 실험에 의해 자동 분할된 데이터와의 비교를 하였다. 비교를 위한 요소는 전문가에 의해 분할된 데이터와 자동 분할된 데이터와의 차이가 16ms, 32ms인 지점의 개수, 전문가에 의해 분할된 지점을 자동 분할 실험에서 분할하지 못한 지점의 개수, 전문가에 의해 분할된 데이터에는 분할하지 않았지만, 자동 분할 실험에서는 분할을 수행한 지점의 개수를 사용하였으며, 이를 이용하여 기존의 실험방법과의 비교를 수행하였다.

5. 실험 결과

표 1은 기존의 방법을 이용한 실험과 본 논문에서 제안한 방법을 사용한 실험 결과로서 표에 나온 파라메터들을 적용하여 각기 다르게 실험 한 후 손으로 분할된 데이터와 비교를 한 결과이다. In 16ms, In 32ms는 각각 오차가 16ms, 32ms이내로 차이 나는 분할한 지점의 개수를 나타내고, lost 는 실제 분할해야 하는 지점을 분할하지 못한 개수, over는 분할하지 않아야 할 부분을 분할한 지점의 개수를 나타낸다.

Threshold의 증가에 따라 16ms안에 들어오는 segment가 늘고, lost segment가 줄어드는 대신 over segment가 늘어나는 것을 볼 수 있다. 이는 threshold를 크게 하면, 음성신호의 변화가 작은 부분들도 분할되기 때문에 실제 작은 변화를 나타내는 음소를 찾아낼 수 있는 반면, 같은 음소지만 발화도중 약간의 떨림과 같은 부분에도 민감하게 반응함으로써 생기는 결과이다. 본 실험에서 제안한 방법은 threshold에 기본 방법에 비해 덜 민감한 것을 볼 수 있다. 이는 decision rule에서 전체 음성의 프레임간 유사도에서 local minimum을 찾기 때문에, 유사도가 threshold보다 작다고 해서 분할 되는 것이 아니라, 지역적으로 가장 작은 값에서만 분할되기 때문이다.

그림 2와 그림 3은 각각 본 실험에서 제안한

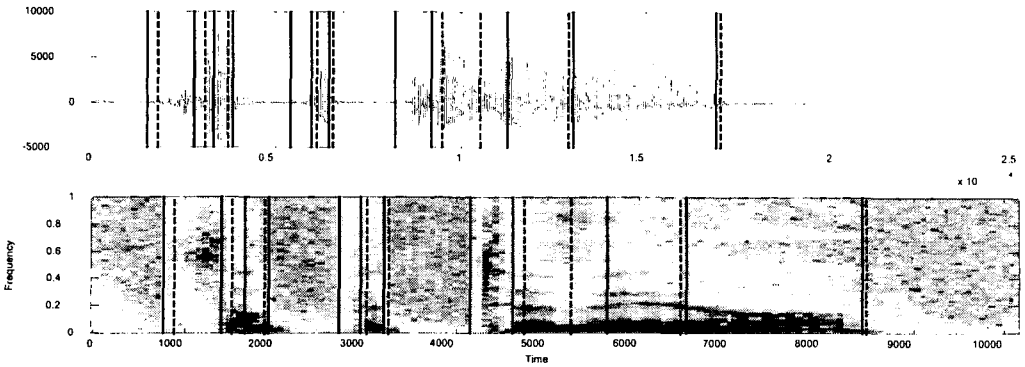

그림 2 제안한 방법에 의해 분할된 결과 (대상음성 :'소프트웨어', 실선: 실험에 의해 분할된 결과, 점선: 전문가에 의해 손으로 분할한 결과)

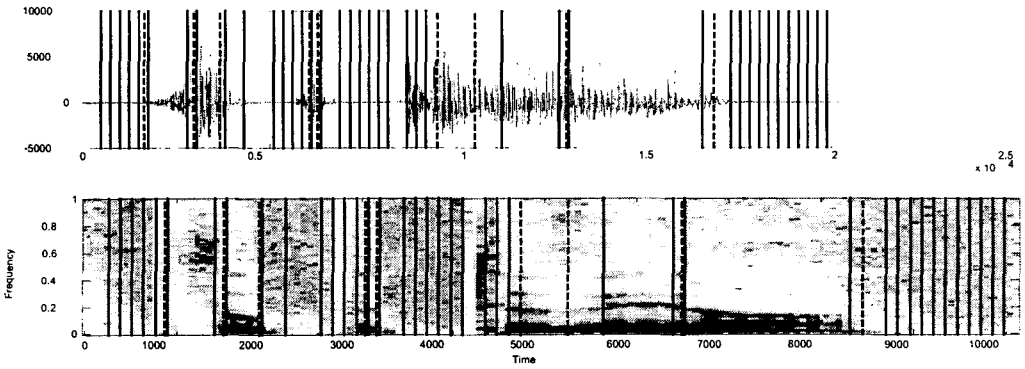

그림 3 기존의 방법에 의해 분할된 결과 (대상음성 :'소프트웨어', 실선: 실험에 의해 분할된 결과, 점선: 전문가에 의해 손으로 분할한 결과)

방법과 기존 방법을 사용하여 '소프트웨어'라는 발음을 분할 한 결과를 음성파형과, 스펙트로그램으로 나타낸 그림이다. 각각 실선은 실험을 통해 자동으로 분할된 결과를 나타내고, 점선은 음성전문가에 의해 분할된 결과를 나타낸다. 그림 2와 그림 3에서 스펙트럼 성분이 크게 변하는 부분에서 분할이 되는 것을 볼 수 있다. 또한 그림 3에서 무성음 구간과 음성 구간 내에서 많은 over segmentation point를 볼 수 있다.

기존 방법 실험에서는 전체적인 정확도가 낮고, 특히 무음구간에서의 over segmentation point가 많은 것을 볼 수 있다. 그 이유는 무음 구간에서는 불규칙한 잡음 성분만 포함되어 있으므로, 프레임간의 유사도가 매우 작기 때문에 나타나는 현상이다.

6. 결론

본 실험은 음성신호를 프레임별로 나눈 후 각 프레임간 스펙트럼 정보의 유사도를 측정하여 한국어 음성 데이터베이스의 음소단위 분할을 목표로 실험하였다. 실험을 통하여, 스펙트럼 정보의 유사도 측정만으로 84%를 넘는 음소분할결과를 얻을 수 있었다. 실험결과를 이용하여 음성인식에 사용되는 훈련용 데이터의 자동 음소분할이나 음성 인식기의 인식률 향상, 화자 인증 시스템에 적용하여 성능향상에 도움을 줄 수 있을것으로 예상되며, 립싱크를 위한 음소 분할처리 부분에 이용할 수 있을 것으로 예상된다. 앞으로의 연구는 음소 인식을 통해 over segmention된 구간을 제거하고, 유사한 특성을 보이는 음소구간에서의 분할을 강화하는 방법에 대한 연구를 해야 할 필요성이 있다.

감사의 글

본 논문은 정보통신부에서 지원하는 대학기초연구지원사업(과제번호: 2001-088-2)으로 수행되었습니다.

참고문헌

[1] Ming-Tzaw Lin, Ching-Kuen Lee and Chin-Yi Lin, "Consonant/vowel segmentation for Mandarin syllable recognition," Computer Speech and Language, Vol. 13, No. 3, pp. 207-222, 1999

[2] Witold Pedrycz, "Fuzzy Multimodels," IEEE Trans. on fuzzy systems. Vol. 4, No. 2, May 1996

[3] Regine Andre-obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," IEEE Trans. On acoustics, speech and signal processing, Vol. 36, No. 1, January 1988

[4] Xuedong Huang, Alejandro Acero and Hsiao-Wuen Hon, *Spoken language processing*, Prentice-Hall International, Inc., pp.427-428, 2001