2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

DAC(Divide-And-Conquer) 기반 분할 알고리즘

구찬모, 왕지남
아주대학교 산업공학과
e-mail:{chanmo, gnwang}@madang.ajou.ac.kr

DAC(Divide-And-Conquer) Based Segmentation Algorithm

Chan-Mo Koo, Gi-Nam Wang
Dept of Industrial Engineering, Ajou University

요약

본 논문은 음운 및 음향학적인 정보를 최대한 이용하고 분할에러를 줄이기 위해서 조절 메카니즘의 하나로 DAC(Divide And Conquer)개념을 사용하여 음성을 *speechlet*으로 나누고(signal localization) 나누어진 음성구간에 대해서 레이블링을 시도(case study)하는 *DAC기반 분할알고리즘*을 제안한다. HMM과 같은 통계학적인 방법을 이용하지 않고 음운학적, 음향학적 지식만을 이용하는 신뢰할 수 있는 분할 알고리즘이며 대용량 음성DB에 대한 레이블링 작업을 단시간에 수행할 수 있고 일관성이 있으며 효과적인 음성엔진 구현 및 음성합성, 화자인증에도 이용 가치가 높다.

1. 서 론

대부분의 레이블링 작업은 음성신호의 파형을 보여주는 툴을 이용한 사람에 의한 레이블링이었다. 따라서, 레이블링 전략을 찾기가 어렵고 사람의 음향학적, 시각적 능력의 차로 인해서 사람에 의한 레이블링은 일관성이 많이 결여되어 있고 많은 시간이 소요된다.

많은 ASL(Automatic Segmentation and Labeling) 시스템이 개발이 되어 레이블링 시간을 상당히 줄였고, 일관된 레이블링을 수행한다. 하지만, 불행하게도 신뢰성에는 한계가 있다. 대부분의 ASL시스템은 HMM(Hidden Markov Model)과 같은 통계학적인 패턴인식접근법을 이용하고 있으며 사람에 의한 수정작업을 병행하기도 한다[2, 4].

본 논문에서는 음절이 잘 발달되어 있는 한국어에 대해서 신뢰할 수 있는 완전히 자동화된 레이블링 시스템을 제안한다. 음소의 음운학적, 음향학적인 정보를 이용하며 음소의 발음표기가 주어지면 음성구간을 찾아내고 음소의 경계를 찾는다. 음소의 경계를 찾는데 유사성을 측정하는 기법중의 하나인 SVF (Spectral Variation Function)을 이용한다[1].

효과적인 조절 메카니즘으로서 DAC방법을 제안한다. 먼저, 음성신호를 *speechlet*이라고 불리우는 여러 개의 지역적 신호들로 나눈다. 이 신호들은 몇 가지 음운학적 사례를 포함하는 정의된 음향학적 패턴(acoustic patterns) 중의 하나와 대응된다. 각각의 음운학적 사례는 음향학적 지식을 포함하고 있으며 각 *speechlet*의 초기 음운학적 경계를 제공한다. 최종 음운학적 경계는 초기 음운학적 경계의 이웃들의 유사성 측정(SVF)을 통해서 결정된다.

본 논문은 HMM과 같은 통계학적인 방법을 이용하지 않고 음운학적, 음향학적 지식만을 이용하는 신뢰할 수 있는 DAC 기반 분할 알고리즘을 제안한다.

2. 음운학적 가정

ASL 시스템에서 FBANK(Filter Bank), LPC(Linear Predictive Coding), 에너지, 켑스트럼 같은 다양한 스펙트럴 파라메터가 이용되는데 본 논문에서는 FBANK를 이용하였다. 먼저, 선택된 스펙트럴 파라메터에 대해서 신뢰할 수 있는 ASL 시스템을 디자인하기 위한 음향학적인 선결요건(가정)에 대해서 고려해보자.

음향학적인 선결요건(가정)에 대해서 MLS (Multi-level segmentation)를 개발한 Glass[3]의 의견을 수용했다.

(A1) 음성은 반 안정적(quasi-stationary) 음성 신호들의 시간적 연속이다.

(A2) 음향학적 신호는 지역적 환경에 영향을 받아 변화되는 짧은 사건이다.

(A3) 음성 신호는 음운학적, 음향학적인 정보를 이용해서 검출 가능한 부분들의 결합이다. 예를 들어, 음절이 발달된 언어에서 모음과 자음은 쉽게 구별이 된다.

가정 (A1)은 같은 분할영역내에 있는 음성 벡터는 다른 분할영역내에 있는 음성 벡터들 보다 더 서로간에 유사하다. 따라서, 분할 문제는 근접한 프레임의 유사성에 의존하는 지역적 클러스터링문제로 줄어들게 된다. 유사성 측정의 수단으로 모든 프레임들은 이전의 그리고 이후의 프레임들과 비교되어진다.

가정 (A2)의 결과로 음성 DATA는 과분할

(over-segment) 혹은 미분할(under-segment) 될 수 있다. 하나의 음소를 가진 음성신호가 여러 개로 분할될 수 있고, 여러 개의 음소를 가지지만 하나의 음소로 분할될 수도 있다. 과분할 혹은 미분할 문제를 해결하기 위해서 두 개의 근접한 지역을 합치거나 하나의 지역을 두 개 혹은 그 이상으로 나누어주는 메카니즘으로 DAC 방법을 제안한다.

가정 (A3)은 본 논문에서 새롭게 사용한 것이며 여전히 논쟁의 여지가 있고 음절이 발달된 언어에 의존적이다. 하지만, 대부분의 경우 모음은 에너지 레벨 혹은 고주파와 저주파 에너지의 비율의 지역 극대치와 관련이 크다. 일부 무성 자음의 경우 에너지 레벨에서 지역 극대치가 있지만 무성 자음은 고주파 에너지만 크고 저주파 에너지는 아주 작으며 모음은 포만트를 형성하지만 무성자음은 고주파 에너지만 높을 뿐이기 때문에 모음과 잡음은 음운학적, 음향학적 정보로부터 찾을 수 있다. 우선 모음의 중간점과 잡음구간을 찾아내고 이를 분할 점으로 사용하여 신호 지역화(signal localization)을 수행한다.

3. DAC 기반 자동 분할 알고리즘

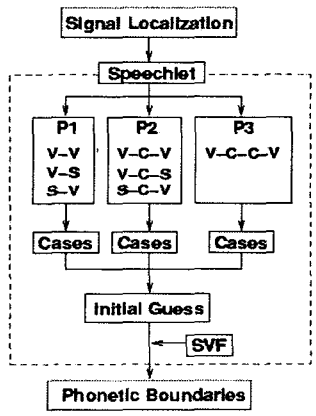

FIG 1 : ASL Process for Korean

음성신호에 대한 DAC 알고리즘은 다음과 같이 정의 할 수 있다.

DAC algorithm: signal localization and case studies for localized signal pieces.

그리고, 한국어의 음절은 다음중의 하나에 속한다.

V, C-V, V-C, and C-V-C (C=자음, V=모음)

음성신호가 모음과 자음에 의해서 지역화 되면, 각각의 지역화된 신호의 음소들은 다음 패턴중의 하나가 된다.

p_1 : V-V, V-S, S-V

p_2 : V-C-V, V-C-S, S-C-V

p_3 : V-C-C-V

이 지역화된 신호 조각을 *speechlet*이라고 부른다. 패턴이 p_k인 한 *speechlet*에서 k개의 음소 경계를 찾으면 된다. 그림 1을 보면 쉽게 이해 할 수 있다.

일단 잡음구간과 모음을 정확하게 구분하고 나면 각 *speechlet*에 대한 구체적인 분할 전략을 적용시킨다. 예를 들어, 패턴 V-C-S에 대해서 C와 S의 경계는 에너지 레벨을 사용해서 쉽게 구별할 수 있다. C는 비음, 폐음, 반모음 등의 특성을 가지고 있고, 모음에 대해서도 포만트 구조가 다를 수 있기 때문에 많은 선택가능한 사례가 존재한다. 하지만, 사례연구(case studies)를 통해서 V와 C사이의 음소 경계를 포함하는 구간에 대한 초기 추측을 할 수 있다: 실제 음소 경계는 구간사이의 SVF 중에서 가장 큰 피크를 선택 할 수 있다. 각 *speechlets*에서 음소 경계를 찾는 것이 바로 기존 연구와 다른 차별성이라 할 수 있으며 상당히 분할 에러를 줄여 준다.

4. 신호 지역화(signal localization)

신호 지역화는 전체 음성구간을 잡음구간과 모음의 중간점을 이용하여 *speechlet*으로 나누고 나누어진 지역적 신호들에 대해서 연속적으로 분할을 수행하는 프로세스이다. 잡음구간을 구별해 내는 것은 쉽기 때문에 모음을 찾는 전략에 대해서 자세하게 설명하고자 한다.

대부분의 경우 모음은 에너지 레벨에서 지역적 극대치를 가진다. 하지만 모든 지역 극대치가 모음이라고 할 수는 없다(Over-Shooting). /s/, /ss/, /ch/, /k/, /t/, 그리고, /p/와 같은 무성 자음들은 높은 에너지를 가질 수 있기 때문이다. 원칙적으로 자음에 대해서는 포만트가 나타나지 않는 다는 사실을 이용해서 구별하거나 자음에 대해서 영점교차율이 높다는 것을 이용해서 구별할 수 있다. 모음은 그 에너지가 포만트의 형태로 전 범위의 주파수대에서 에너지가 나타나지만 무성 자음은 고주파 밴드에서만 에너지가 나타난다. 따라서, 고주파 에너지가 상당히 높은 지역 극대치를 무시할 수 있다. 하지만, 모음과 자음을 구별하는 문제는 단순하지 않다. 반모음(semi-vowel: /l/, /r/)과 비음(nasals: /m/, /n/, /ng/)의 에너지 수준 과 모음의 에너지 수준을 구별하는 전략을 수립해야 한다.

모음의 에너지 수준에서 지역 극대치가 나타나지 않을 수가 있다(Under-Shooting). 예를 들어, /d-a eu-m ch-a-ng/(다음창)을 발음할 경우 /eu/ 음에서 지역 극대치가 나타나지 않을 수 있다. 이는 특히, 모음이 연속으로 오고 아주 빠르게 발음이 되는 경우에 발생빈도가 높다. 그리고 이 모음들(특히, a-o, u-eo)은 제 1 포만트와 제 2 포만트에서의 주파수가 유사하다는 것을 알 수 있다. 대부분의 Under - Shooting 문제는 저주파와 고주파의 에너지 비율(energy ratio)을 통해서 해결할 수 있다. 이 에너지 비율은 다음과 같이 정의된다.

$$G_n = \sum_{i=0}^{m/2-1} S_n(i)^2 / \sum_{i=m/2}^{m-1} S_n(i)^2 \quad (5)$$

이때 m은 음성 벡터, S_n은 영역이다.

5. 사례연구(Case Study)

사례연구는 각 *speechlet*의 시작점과 끝점이 결정된 후 음운학적, 음향학적 지식을 이용해서 *speechlet*의 초기 경계를 결정하는 것을 말한다. 총 패턴의 수는 3가지이며 패턴에 따라 결정하여야할 각 *speechlet*의 초기 경계의 수도 자동 결정된다. 각 케이스(case)별로 초기경계 결정방법에 대해서 자세히 설명하고자 한다.

5.1) P1(경계 수: 1)

5.1.1) VV

모음검출에서 에너지와 고주파와 저주파의 에너지 비율을 이용해서 지역적 최대치를 찾았고 이를 *speechlet*의 경계로 사용하였다. 따라서, 상대적으로 모음과 모음이 연속으로 오는 경우 그 사이에서 지력적 최소가 존재한다. 그러므로, 지역적 최소에서 바로 모음과 모음의 경계가 존재한다라고 할 수 있다. 만약, 지역적 최소가 여러 개 존재하는 경우는 가장 최소 에너지를 가지는 지역적 최소를 초기 경계로 한다.

5.1.2) SV

음성신호의 시작점과 끝점 추출에 의해서 구해진 음성의 시작점을 그 초기 경계로 한다. 경계를 구하기 쉽고 정확하다.

5.1.3) VS

음성신호의 시작점과 끝점 추출에 의해서 구해진 음성의 끝점을 그 초기 경계로 한다. 경계를 구하기 쉽고 정확하다.

5.2) P2(경계 수: 2)

5.2.1) VCV

/ng/을 제외한 종성자음은 다음에 오는 모음에 의해서 영향을 받기 때문에[6] VCV 경우에서 /ng/을 제외한 모든 자음은 초성으로 발음이 된다. 따라서, 대부분의 경우 V-C 사이에서 최소 에너지를 가지게 된다.

자음이 비음인 /n/,/m/이면 V-C 사이에서 최소 에너지를 가지는 점을 첫 번째 초기 경계로 한다. 두 번째 초기 경계는 C-V사이의 지역적 최소 에너지를 가지는 지점으로 한다. 왜냐하면, /n/,/m/은 비음인 동시에 유성자음이기 때문에 에너지가 존재하며 C-V의 경계에서 지역적 최소 에너지를 가지기 때문이다.

비음 /ng/인 자음이 오는 경우는 두 번째 초기경계 지점에서 최소 에너지를 가지게 된다. 비음 /ng/가 종성이기 때문이다. 첫 번째 경계는 V-C 사이의 지역적 최소 에너지를 가지는 점을 이용한다. 지역적 최소가 존재하지 않는다면 큰 peak을 찾거나 두 번째 경계에서 threshold값을 적용하여 결정한다.

자음이 파열음인 경우 고주파 에너지 레벨이 높다는 음운학적 지식을 이용한다. 또한 파열음을 발음하기 위해서 *short pause* 현상이 일어난다. 따라서, 최소 에너지를 가지는 점을 첫 번째 경계로 하고 두 번째 경계는 고주파 에너지를 가지는 점에서부터 최소 에너지를 가지는 점을 그 경계로 한다.

자음이 /r/, /h/인 경우는 지역적 최소 에너지를 갖는 프레임을 첫 번째 초기 경계로 하고 이점을 기준으로 2~3 프레임의 threshold값을 적용해서 두 번째 초기 경계를 결정한다.

5.2.2) VCS

자음은 폐쇄음이거나 비음인 종성으로 쓰이는 자음밖에 올 수가 없다.

무성음이고 폐쇄음인 자음의 경우는 대부분 8 프레임을 초과하지 않는다. 따라서, 두 번째 경계점에서 8프레임정도 뒤쪽의 피크를 첫 번째 경계로 한다. 두 번째 경계점은 음성의 끝점을 그 초기 경계로 한다.

비음인 자음(/n/, /l/, /m/)이 오는 경우는 이미 결정된 두 번째 경계로부터 지역적 최소 에너지를 가지는 점을 찾는다. 만약, 최소 에너지가 존재하지 않는 경우 두 번째 경계로부터 6~8 프레임의 threshold값을 적용해서 첫 번째 초기 경계를 구한다.

5.2.3) SCV

종성으로 쓰이는 자음은 절대로 올 수 없다.

묵음 다음에 비음인 자음이 오는 경우 G(에너지 Ratio)가 가장 작아지는 지점을 두 번째 경계로 한다. 첫 번째 경계는 음성의 시작점을 그 경계로 한다.

묵음 다음의 자음이 무성음이고 파열음인 경우 고주파수 레벨의 에너지가 높다. 모음 또한 파열 자음만큼은 아니지만 상당히 높은 에너지를 보이므로 자음에서 모음으로 변화하는 지점에서 상대적으로 에너지가 낮아진다. 따라서, 고주파 레벨 에너지를 구해서 고주파 에너지가 가장 작아지는 점을 찾아 두 번째 경계로 한다.

5.3) P3(경계 수: 3)

5.3.1) VCCV

10.3.1) VCCV

첫 번째 자음은 종성 자음이고 두 번째 자음에는 종성으로 쓰이는 자음은 올 수가 없다. 이때 종성자음은 비음인 유성음 혹은 폐쇄음인 무성자음일 수 있다. 그리고, 종성으로 쓰이지 않는 자음은 비음인 유성음 혹은 파열음이거나 /r/,/h/인 무성자음일 수 있다. 따라서, 다음 4가지의 경우가 존재한다.

10.3.1.1) V-VC-UC-V

두 번째 자음인 UC(Voiced Consonant)를 발음하기 위해서 short pause 현상이 일어남으로 자음과 자음 사이에서 최소 에너지를 갖는다. 따라서, 최소 에너지를 가지는 지점을 두 번째 경계로 정하고 첫 번째 경계는 V(Vowel)-VC(Unvoiced Consonant) 사이에서 지역적 최소 에너지를 가지는 지점을 찾아 초기 경계로 이용한다.

두 번째 자음이 무성음이고 파열음인 경우 고주파수 레벨의 에너지가 높다. 그리고, 모음 또한 파열 자음만큼은 아니지만 상당히 높은 고주파 에너지를 보이므로 자음에서 모음으로 변화하는 지점에서 상대적으로 고주파 에너지가 낮아진다. 따라서, 고주파 레벨 에너지를 구해서 에너지가 가장 작아지는 점을 찾아 세 번째 경계로 한다.

두 번째 자음이 /r/,/h/인 경우는 두 번째 경계에서

2~3 프레임정도의 threshold값을 적용해서 세 번째 경계를 결정한다.

10.3.1.2) V-UC-VC-V

UC에는 폐쇄음인 종성자음이 온다. 그리고 VC로 올 수 있는 자음은 /n/, /m/ 뿐이다. 따라서, 폐쇄음을 발음한 이후 극히 짧은 묵음 구간이 발생하며 최소 에너지를 가지는 지점이 바로 UC-VC의 경계가 된다. 첫 번째 경계는 무성음이고 폐쇄음인 자음의 경우는 대부분 8 프레임을 초과하지 않는다는 사실을 이용하여 두 번째 경계 점에서 8프레임정도 뒤쪽의 피크를 첫 번째 경계로 한다. 세 번째 경계는 VC-V 사이에서 지역적 최소 에너지를 가지는 지점을 찾아 초기 경계로 이용한다.

10.3.1.3) V-UC-UC-V

첫 번째 UC에는 폐쇄음인 종성자음이 온다. 그리고, 두 번째 UC는 파열음, 혹은 /r/, /h/음이다. 폐쇄음을 발음한 이후 극히 짧은 묵음 구간이 발생하므로 이 지점에서 최소 에너지가 존재하게 되고 두 번째 경계로 사용된다. 첫 번째 경계는 무성음이고 폐쇄음인 자음의 경우는 대부분 8 프레임을 초과하지 않는다는 것을 이용한다. 따라서, 두 번째 경계 점에서 8프레임정도 뒤쪽의 피크를 첫 번째 경계로 한다.

두 번째 자음이 무성음이고 파열음인 경우 고주파수 레벨의 에너지가 높다. 그리고, 모음 또한 파열자음만큼은 아니지만 높은 고주파 에너지를 보이므로 자음에서 모음으로 변화하는 지점에서 상대적으로 낮은 고주파 에너지가 존재한다. 따라서, 고주파 레벨 에너지를 구해서 에너지가 가장 작아지는 점을 찾아 세 번째 경계로 한다.

두 번째 자음이 /r/,/h/인 경우는 두 번째 경계에서 2~3 프레임정도의 threshold값을 적용해서 세 번째 경계를 결정한다.

10.3.1.4) V-VC-VC-V

두 자음 모두 유성 자음으로 이루어진 경우로 두 번째 VC에는 /n/,/m/만 올 수 있다. VC-VC 사이에서 지역적 최소 에너지가 존재한다면 그 점을 두 번째 경계로 한다.V-VC, VC-V의 경계는 지역적 최소 에너지를 가지는 지점을 찾아 그 첫 번째와 세 번째 경계로 이용한다.

구분	In 20ms		In 30ms		In 40ms		Total boundaries
S-V	160	97.56%	162	98.78%	162	98.78%	164
S-C-V	1412	91.81%	1480	96.23%	1507	97.98%	1538
V-C-C-V	427	75.31%	462	81.48%	492	86.77%	567
V-C-V	982	83.93%	1067	91.20%	1109	94.79%	1170
V-S	483	95.08%	487	95.87%	494	97.24%	508
V-C-S	707	83.97%	758	90.02%	786	93.35%	842
V-V	172	68.25%	196	77.78%	216	85.71%	252
Total	4343	86.15%	4612	91.49%	4766	94.54%	5041

그림 2 Accuracy of Segmentation Algorithm

5. 실험결과

본 자동 음소 레이블링 알고리즘은 원광대학교 이용주 교수 팀에서 개발한 PRW 3813종 40set(남녀 총 500명분) 중에서 남녀 각각 5명씩, 총 900여 단어를 대상으로 실험을 수행하였다. 새로운 알고리즘의 성능을 평가하기 위해서 음운학적 지식이 있는 사람에 의한 레이블링을 수행하였으며, 20ms 내에서 86.15%의 정확성을 보였다.

5. 결 론

음성신호에 대해서 음운학적, 음향학적 정보를 이용하는 DAC(Divide-and-Conquer)기반의 자동 레이블링 알고리즘을 소개하였다. 이 자동 레이블링 알고리즘은 현재 20ms이내에서 86%정도의 정확성을 보이고 있다.

차후 연구로서 모음과 모음이 연속해서 오는 경우 특히, 이중모음이 오는 경우에 대한 연구가 더 진행되어야 하며 자음에 대해서도 더욱더 다양한 전략이 필요하다. 특히, 모음 다음에 유성자음(/l/,/m/,/n/,/ng/)이 오는 경우 그 경계를 구분하기가 어렵고, VCCV case에서 폐쇄음인 자음 다음에 /s/,/ss/ 자음이 오는 경우에는 폐쇄음 다음에 short pause가 아니올 수 있다. 이러한 문제점들에 대한 성능 개선을 위해서 많은 음성DB에 대해 자동레이블링 알고리즘의 성능을 평가하고 새로운 전략을 수립하여야 한다.

본 DAC기반 자동분할 알고리즘은 HMM 방식과는 달리 훈련과정이 필요 없으며 대용량 음성DB에 대한 labeling 작업을 단시간에 수행할 수 있고 일관성이 있다. 효과적인 음성엔진 구현 및 음성합성 나아가 화자인식에도 이용 가치가 높다.

참고문헌

[1] F. Brugnara, D. Falavigna, and M. Omologo, *Automatic segmentation and labelling of speech based on hidden Markov models*, Speech Communication 12 (1993), 357-370.

[2] S. Cox, R. Brady, and P. Jackson, *Techniques for accurate automatic annotation of speech wave forms*, Proceedings of ICSLP'98 (Sydney, Australia), December 1998, pp. 1947-1950.

[3] J.R. Glass, *Finding acoustic regularities in speech: Application to phonetic recognition*, Ph.D. Thesis, MIT Press, May 1988.

[4] J.-P. Hosom, *Automatic time alignment of phonemes using acoustic-phonetic information*, Ph.d. thesis, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, May 2000.

[5] K. Kvale, *On the connection between mannual segmentation conventions and error made by automatic segmentation*, Proceedings of ICSLP'94 (Yokohama, Japan), September 1994, pp. 1667-1670.

[6] 허웅, "국어음운학", 정음사, 1984, pp. 129-280.0.