자동 구두점 삽입을 이용한 Rich Transcription 생성

김지환(LG전자기술원)

<차 례>

<Abstract>

Rich Transcription Generation Using Automatic Insertion of Punctuation Marks

Ji-Hwan Kim

A punctuation generation system which combines prosodic information with acoustic and language model information is presented. Experiments have been conducted first for the reference text transcriptions. In these experiments, prosodic information was shown to be more useful than language model information. When these information sources are combined, an F-measure of up to 0.7830 was obtained for adding punctuation to a reference transcription.

This method of punctuation generation can also be applied to the 1-best output of a speech recogniser. The 1-best output is first time aligned. Based on the time alignment information, prosodic features are generated. As in the approach applied in the punctuation generation for reference transcriptions, the best sequence of punctuation marks for this 1-best output is found using the prosodic feature model and an language model trained on texts which contain punctuation marks.

* Keywords: Punctuation generation, Prosody, Classification and regression tree(CART)

1. 서 론

음성으로부터의 정보 검색(information retrieval)은 음성인식(speech recognition) 기술이 언어이해(speech understanding)의 수준으로 발전함에 있어서 가장 중요한 단계들 중 하나이다. 현재 음성을 포함하는 멀티미디어 자료에 대한 효과적인 검색 방법에 대한 연구가 활발히 진행 중이며, 특히 검색의 결과로서 제공해야 하는 멀티미디어 자료의 시작 지점과 끝 지점을 판단하는 연구가 활발히 진행되고 있다. 일반적으로 검색의 결과는 문장 단위로 구성된다. 문장의 경계에는 많은 경우 구두점이 위치하게 되므로 정확한 자동 구두점의 생성은 효과적인 정보 검색을 위해서 가장 필수적인 연구가 된다.

자동 구두점 생성에 관한 연구의 중요성에 대한 다른 이유는 구두점이 자동으로 생성됨으로서 음성인식기의 출력, 즉, 발성에 대한 전사(transcription)의 가독성이 크게 향상되기 때문이다. 어휘(vocabulary)에 있는 단어들로 조합되는 단어열 중, 확률이 가장 높은 단어열이 음성인식 결과로 출력되기 때문에, 음성인식 오류가 없는 경우에도 Standard Normalised Orthographical Representation(SNOR)로 알려져 있는 표준 음성인식기 출력 포맷은 구두점과 숫자를 포함하지 않게 된다[1].

따라서, 정확한 구두점이 생성된다면 음성인식기의 음성인식 결과의 가독성은 크게 향상되게 된다. 받아쓰기(dictation)가 수행되는 경우, 받아쓰기 시스템은 화자에게 필요한 시점에 "마침표", "쉼표" 등을 발성하게 함으로서 구두점을 생성할 수 있다. 그러나 뉴스 자료에 대한 음성인식과 같이 화자가 자신의 발성이 자동으로 음성인식 되고 있다는 모르는 경우에는, 이와 같은 발성된 구두점 정보는 얻을 수가 없다. 특히 음성 자료의 입력의 경우 음성인식 오류로 인하여 자동 구두점 생성은 더욱 어려운 문제가 되게 된다.

본 논문에서는 음성 자료에 대한 전사가 주어진 경우, 이에 대해 prosody 정보와 구두점을 포함한 확장된 언어모델을 이용하여 자동으로 구두점을 생성하는 방법을 제시하고, 이 방법을 기준 전사(reference transcription)와 음성인식기의 1-best 결과에 대해서 검증한다.

본 논문은 6개의 장으로 구성되어 있다. 2장에서는 기존의 연구를 소개한다. 3장에서는 운율을 이용한 구두점 생성 방법을 제시한다. 4장에서는 뉴스자료를 이용한 실험 방법과 결과 측정 방법을 설명한다. 5장에서는 기준 전사에 대한 구두점 생성의 결과를 보여주며 또한 같은 방법이 음성인식기의 1-best 결과에 적용되었을 경우의 결과도 보여준다. 마지막으로 6장에서는 본 논문에 대한 결론을 내린다.

---

1) 영어 음성인식의 경우 SNOR이 대소문자 구분을 하지 못하는 것도 가독성을 떨어뜨리는 또 다른 주요 이유 중 하나임.

2. 기존의 연구

Beeferman, Berger, Lafferty는 lexical 정보에 기반한 자동 구두점 생성 시스템으로 Cyberpunc라 불리는 시스템을 개발했다[1]. Cyberpunc는 문장의 경계가 미리 주어졌다는 가정하에 쉼표만을 생성했다. 구두점이 없는 문장의 후처리 과정에서 구두점을 고려한 확장된 언어모델을 적용하여 쉼표를 생성한다.

Chen은 acoustic 정보와 lexical 정보에 기반한 구두점 생성을 포함한 음성인식 방법을 제안하고, 3명 화자의 낭독체 음성자료(read speech)를 이용하여 제안된 방법을 검증했다[2]. 이 연구에서는 구두점들이 음성인식기에서 개별 단어로서 다루어졌다. 따라서, 음성인식과 동시에 구두점 생성이 수행될 때 각각의 구두점에 대해서 발음 사전(pronunciation dictionary)에 등록하는 발음을 정해야 하는데, 구두점에 대한 발음으로 silence, breath 또는 다른 non-speech sound들을 등록했다.

많은 수의 마침표와 의문 부호가 문장의 마지막에 위치하기 때문에, 문장의 경계를 정확히 인식하는 것은 구두점 생성에서 매우 중요하다. Gotoh와 Renals는 lexical 정보와 pause duration을 이용한 문장 경계 인식기를 개발했다[3]. 문장 경계 인식은 각 음성인식 결과의 단어열에 대해서 해당하는 문장 경계 클래스, 즉 "문장 끝" 클래스 또는 "문장의 끝이 아님" 클래스의 열을 찾도록 개발되었는데, 언어모델과 pause duration 모델로부터의 확률 값의 결합에 의해서 문장 경계 인식이 수행되었다.

문장의 끝에 위치한 구두점의 경우, 동일한 구두점도 다른 형태로 사용되어 질 수 있다. 예를 들어, 마침표는 축약형을 위해서도 사용가능하고, 소수점 표시, 문장의 끝을 나타내는 표시, 또는 문장 끝에 위치한 축약형의 표시를 위해서도 사용되어 진다. Palmer와 Hearst는 동일 구두점에 대해서 구두점의 타입을 분류하는 학습 가능한 시스템을 소개했다[4]. 이 시스템에서는 구두점 주위의 단어의 품사가 예측되어졌고, 예측된 품사를 바탕으로 구두점의 타입이 분류되었다.

Shriberg와 Bates 등은 discourse 구조와 prosodic 정보간의 강한 상관관계를 검증했다[5]. Fach는 문법적인 phrasing과 prosodic phrasing간의 비교를 수행했다[6]. 이 연구에서 문법적 구조는 Abney chunk parser[7]를 이용해서 생성했고, prosodic 구조는 ToBI[8] 레이블 파일로 주어졌다. 이 연구에서 낭독체 음성 자료에 대해 최소한 65% 문법적 경계는 prosodic 경계로서 표현이 된다는 것이 밝혀졌다.

Taylor와 King등은 대화체 음성(spontaneous dialogue speech)에 대한 음성인식에서 intonation을 이용해서 단어 인식률을 줄이는 방법을 기술했다[9]. 이 연구에서 각각의 Dialogue Act(DA. 발성의 타입에 대한 구분. 예: 평서문, 의문문, 동의 등)에 대해서 별도의 intonation 모델이 적용되어져서 발성에 대한 DA열이 생성되고, 각각의 DA 타입에 대해 별도의 언어모델이 적용되어 새로운 음성인식 결과를 생성한다.

자동 구두점 생성을 포함한 discourse 구조 분석에서 prosodic 정보를 이용하기 위해서는 prosodic feature 값을 얻는 방법과, prosodic feature 모델의 생성 방법, 그리고 다른 정보에 대한 모델과 prosodic feature 모델을 결합하는 방법에 대해서 많은 주의가 기울여 져야 한다.

Shriberg와 Bates 등은 단어 모델과 prosodic feature 모델의 결합 방법을 논의했다[5]. 이 연구에서 두 모델의 결합 방법은 DA 식별에 적용이 되었다. Classification And Regression Tree(CART)[10]가 prosodic feature 모델을 생성하는데 사용되었다. 연산량을 tractable하게 만들기 위해서 prosodic feature가 DA에 한번 condition이 되면, 단어에는 independent하다는 가정을 허용했다[2].

3. 자동 구두점 생성

본 장에서는 prosody 정보와 구두점을 포함한 확장된 언어모델을 이용하여 자동으로 구두점을 생성하는 방법을 기술한다. 기준 전사에 대해서 자동 구두점 생성이 수행될 때, 단어열은 이미 주어진 상태이다. 따라서, 실험은 단어간의 구두점을 삽입하는데 초점이 맞추어 진다.

Y를 구두점 기호 열이라고 하고, W를 단어열, F를 해당하는 prosodic feature 열이라고 하자. 자동 구두점 생성 시스템은 W와 F가 주어졌을 때 식 (1)을 최대로 하는 maximum *a posteriori* Y인 Y_{MAP}를 찾는 것이다.

$$Y_{MAP} = \arg_{Y} \max P(Y|W,F) \tag{1}$$

$P(Y|W,F)$에 대해서 수식을 전개하면,

$$\begin{aligned} P(Y|W,F) &= \frac{P(Y,W,F)}{P(W,F)} = \frac{P(Y,W,F)}{P(W,F)} \frac{P(Y,W)}{P(Y,W)} \frac{P(W)}{P(W)} \\ &= \frac{P(Y,W,F)}{P(W,F)} \frac{P(Y,W)}{P(W)} \frac{P(W)}{P(Y,W)} \\ &= \frac{P(F|Y,W)P(Y|W)}{P(F|W)} \end{aligned} \tag{2}$$

Y가 $P(F|W)$에 독립이기 때문에

2) 비슷한 가정은 Taylor와 King 등의 연구[9]에서 도입되었었다.

$$P(Y|W,F) \propto P(F|Y,W)P(Y|W) \tag{3}$$

F가 Y에만 종속이라고 가정하고, $P(F)$가 uniformly distributed하다고 가정하면,

$$P(F|Y,W) = P(F|Y) = \frac{P(Y|F)P(F)}{P(Y)} \propto \frac{P(Y|F)}{P(Y)} \tag{4}$$

따라서,

$$P(Y|W,F) \propto P(F|Y,W)P(Y|W) \propto \frac{P(Y|F)}{P(Y)} P(Y|W) \tag{5}$$

y_i를 i번째 위치의 구두점이라 하고, f_i를 i번째 prosodic feature라 하자. 1st order Markov 가정을 적용하면

$$p(y_i|f_1,\ldots,f_n) \simeq p(y_i|f_i) \tag{6}$$

또한 y_i가 conditionally independent하다고 하면

$$p(y_1,\ldots,y_n|F) = \prod_{i=1}^{n} p(y_i|F) \tag{7}$$

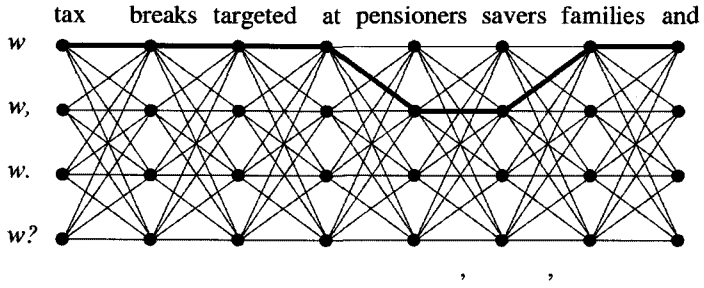
따라서, $P(Y|F)$는

$$P(Y|F) = \prod_{i=1}^{n} p(y_i|f_i) \tag{8}$$

식 (8)에 있는 확률들은 classification tree의 terminal node에서 계산되어 질 수 있다. Classification tree는 4.1장에서 자세히 설명하기로 한다. 식 (5)에서의 $P(Y|W)$는 언어모델로부터 계산되어지고, $P(Y)$는 training data의 count로부터 찾을 수 있다.

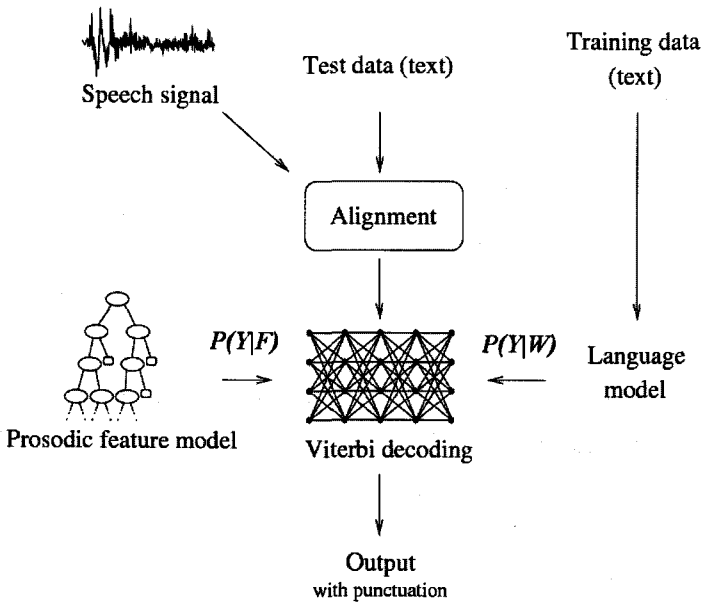
여러 구두점 타입 중, 본 연구는 마침표, 쉼표 그리고 의문부호만을 대상으로 하고 있는데, 이들 세 구두점 타입만이 모델을 생성하기에 충분한 양의 학습 데이터를 가지고 있고, 또한 테스트 시에도 정확하게 결과를 측정할 수 있는 양의 데이터가 확보되기 때문이다.

구두점을 가지고 있지 않은 n개의 단어열 $w_1, \ldots, w_n$에 대해서, 각 단어의 끝에 구두점이 놓일 수 있다. 세 개의 구두점 타입과 구두점 없음(No-Punctuation. 이하 NP)을 고려했을 때 총 4^n개의 가능한 후보(hypothesis)가 발생을 하게 된다. 최적의 구두점 열은 Viterbi search 알고리즘[11]을 이용해서 얻을 수 있다. 이 알고리즘을 이용함으로서, 최적의 구두점을 찾는 것에 소요되는 시간은 단어의 수에 linear하게 된다. <그림 1>은 기준 전사의 한 예에 대해서, 구두점 생성에 대한 Viterbi search가 진행되는 것을 보여준다. 그림 안의 굵은 선은 최적의 후보를 나타낸다. 이 후보 상에서는 쉼표가 "pensioners"와 "savers" 뒤에서 생성된다.



<그림 1> 예제 기준 전사의 구두점 생성에 대한 Viterbi search 프로세스. 굵은선은 최적의 구두점 열을 나타낸다. 최적 구두점 열에 따라 생성된 구두점들이 하단에 나타난다.

<그림 2>는 기준 전사에 대한 구두점 생성에 대한 과정을 보여준다. 음성신호는 기준 전사에 따라 time align된다. 이 time alignment 과정에서 각 단어의 시작 지점과 끝 지점이 파악된다. Prosodic feature가 각 단어의 끝에서 계산되어지고, prosodic feature 모델로부터 $P(Y|F)$가 측정된다. 단어와 구두점들의 열에 대한 확률인 $P(Y|W)$는 통계적 언어모델로부터 계산되어 진다. 최적의 구두점열들은 Viterbi search를 통해서 생성된다.



<그림 2> 기준 전사로부터 구두점을 생성하는 과정의 개념도

4. 실험 방법 및 결과 측정 방법

뉴스 텍스트 자료와 100시간 분량의 1998 Hub-4 방송 뉴스(broadcast news) 자료가 학습 자료로서 사용되었다. 뉴스 텍스트 자료는 (이하 BNText92_97) 1992년에서 1997년 사이에 수집된 1.84억 단어로 구성된 뉴스 텍스트 자료이다. 100시간 분량의 1998 Hub-4 방송 뉴스 자료는 (이하 BNAcoustic98) 뉴스에 대한 acoustic 자료와 이에 대한 전사로 구성되어 있다.

NIST 1998 Hub-4 방송 뉴스 벤치마크 테스트에서 사용된 테스트 자료가 (이하 TestBNAcoustic98) 본 연구에서 테스트 자료로 사용되었다. TestBNAcoustic98은 3시간 분량의 acoustic 자료와 이에 대한 전사로 구성되어 있다. 학습 자료와 테스트 자료는 <표 1>에 정리되어 있다.

<표 1> 학습 및 테스트 자료

자료명	내용	단어수	용도	Acoustic 자료 유무
BNtext92_97	1992_97 뉴스 텍스트	184M	학습자료	무
BNAcoustic98	100시간 분량의 1998 Hub-4 자료	774K	학습자료	유
TestBNAcoustic98	1998 벤치마크 테스트 자료	32K	테스트 자료	유

4-gram 언어모델은 BNText92_97과 BNAcoustic98에 대해서 각기 생성한 언어모델을 perplexity 최소화 방법을 사용해서 interpolation해서 생성했다. 학습 자료로서 acoustic 자료는 BNAcoustic98이 유일하기 때문에 prosodic feature 모델의 구현에는 BNAcoustic98만 사용되었다.

평가 척도로는 F-measure[12]와 Slot Error Rate(SER)[12]의 척도를 이용해서 평가하게 되는데, F-measure를 구하기 위해서 필요한 Precision (P)과 Recall (R), 그리고 F-measure (F)와 SER의 정의는 다음과 같다.

$$P = \frac{\text{정확한 구두점 개수}}{hypothesis\text{의 구두점 개수}} \tag{9}$$

$$R = \frac{\text{정확한 구두점 개수}}{reference\text{의 구두점 개수}} \tag{10}$$

$$F=\frac{RP}{(R+P)/2} \tag{11}$$

$$SER=\frac{\text{부정확한 구두점 개수}}{reference\text{의 구두점 개수}} \tag{12}$$

4.1. Prosodic feature 모델 셋업

기계적인 측정이 쉬운 prosodic feature들이 Dialog Act(DA) 식별에 대해 Shriberg와 Bates등의 연구 [5]에서 검증되었고, 자동 topic segmentation에 대해 Stolcke와 Shriberg등의 논문 [13]에서 검증되었으며, 정보추출(information extraction)에 대해서도 Hakkani-Tur와 Tur등의 연구 [14]에서 검증되었다.

DA 식별에 대한 각 prosodic feature의 contribution을 고려한 후, contribution이 높은 feature들 중 자동 구두점 생성 분야에 유용할 것으로 기대되는 10개의 prosodic feature들을 본 연구에 사용했다. <표 2>는 자동 구두점 생성에서 사용된 10개의 prosodic feature들을 나타낸다.

<표 2> 사용한 10개의 prosodic feature들과 CART에 대한 contribution(feature usage: 테스트 자료에 의해서 feature가 query된 횟수. feature appearance: non-terminal node에서 classifying feature로서 사용되어진 횟수. 50Hz ≦ good F0 ≦ 400Hz)

Feature명	내용	Feature appearance	Feature usage
Pau_Len	단어끝에서의 pause 길이	672	0.5799
Dur_fr_Pau	이전 pause로 부터의 duration	539	0.0230
Avg_F0_L	왼쪽 윈도우에서 good F0들의 평균	342	0.0246
Avg_F0_R	오른쪽 윈도우에서 good F0들의 평균	230	0.0363
Avg_F0_Ratio	Avg_F0_R/Avg_F0_L	261	0.0461
Cnt_F0_L	왼쪽 윈도우에서 good F0의 개수	204	0.0429
Cnt_F0_R	오른쪽 윈도우에서 good F0의 개수	230	0.0176
Eng_L	왼쪽 윈도우에서 RMS energy	203	0.0038
Eng_R	온른쪽 윈도우에서 RMS energy	160	0.0252
Eng_Ratio	Eng_R/Eng_L	239	0.2006

각 단어의 끝은 구두점이 위치할 수 있는 후보가 된다. 따라서 alignment 결과에 따른 각 단어의 종료지점에서 prosodic feature들이 측정된다. 윈도우 길이는 0.2초로 세팅 했다. 왼쪽 윈도우는 단어 끝 바로 이전에 위치하는 윈도우가 되고, 오

른쪽 윈도우는 단어의 끝 바로 다음에 위치하는 윈도우가 된다. "Good" F0은 50 Hz와 400 Hz 사이에 놓이는 F0을 뜻한다.

Prosodic feature 모델은 CART를 이용해서 만들어졌다. Classification tree의 생성을 위한 prosodic feature들은 BNAcoustic98에서 측정되었다. CART 생성 시 cross validation 방법이 사용되었다.

각 feature들의 전체적인 contribution은 'feature usage'를 이용해서 측정이 할 수 있는데, feature usage는 테스트 자료에 의해서 feature가 query된 횟수로서 측정이 된다. 또한 'feature appearance'로도 측정이 가능한데, feature appearance는 non-terminal node에서 classifying feature로서 사용되어진 횟수를 나타낸다. 각 feature에 대한 전체적인 contribution은 <표 2>에 정리되어 있다.

Pau_Len과 Eng_Ratio의 feature usage는 약 78%이다. 이 측정 지표는 tree상에서 feature의 위치를 반영한다. Tree상에서 feature의 위치가 높을수록, feature usage 값은 커지게 된다.

5. 구두점 생성 결과

기준 전사에 대해 구두점을 생성하기 위해 언어모델만 사용한 시스템(LMOnly), prosodic feature 모델만 사용한 시스템(CARTOnly), 그리고 언어모델과 prosodic feature 모델을 함께 사용한 시스템(LM+CART)의 세 가지 시스템을 구현했다. LMOnly는 1.85억 개의 단어로 이루어진 전사(BNText92_97 & BNAcoustic98)로 부터 학습 되었다. 이 전사는 구두점을 포함하고 있기 때문에, 언어모델은 구두점을 포함하고 있지 않은 단어 열에 대해서 구두점의 위치와 타입을 예측할 수 있다. 4-gram 언어모델이 LMOnly에서 사용되었다. CARTOnly는 <표 2>에서 기술된 10개의 prosodic feature를 이용하여 생성했다. 이들 feature들은 100시간 분량의 뉴스자료로 부터 측정되었다(BNAcoustic98).

Prosodic feature 모델에 주어진 weight인 scale factor (α)를 사용해서, prosodic feature 모델과 언어 모델에 대한 상대적인 중요도를 조절 할 수 있다. Scale factor는 아래의 식과 같이 두 모델간의 병합에 포함되어진다.

$$\alpha \times \log P(F|Y, W) + \log P(Y|W) \tag{13}$$

언어모델만 사용한 시스템(LMOnly)은 F-measure로 0.5717, SER로 72.25%의 결과를 보였다. LMOnly가 기준 전사에 대해서 구두점을 생성할 때, precision (0.5966)이 recall(0.5488)에 비해서 높게 나왔다. 놀랍게도, prosodic feature만 사용한 시스템(CARTOnly)가 LMOnly에 비해서 성능이 좋게 나왔는데, F-measure로는

0.0521, SER로는 0.54%가 좋게 나왔다. CARTOnly에 대해서는 recall(0.7417)이 precision(0.5383)보다 높게 측정되었다. 이 결과는 CARTOnly가 상대적으로 많은 수의 구두점을 생성하지만, 생성된 구두점의 정확도는 상대적으로 떨어지는 것을 보여준다.

LMOnly는 lexical 정보로부터 구두점을 생성하고, CARTOnly는 또 다른 정보원인 prosodic feature 정보로부터 구두점을 생성하므로 두 개의 모델은 상호보완적일 것이 예측된다. 이는 실험 결과로 부터도 예측이 되는데, CARTOnly에 대해서는 recall이 precision에 비해서 크게 높고, LMOnly에 대해서는 precision이 recall에 비해서 약간 높기 때문이다.

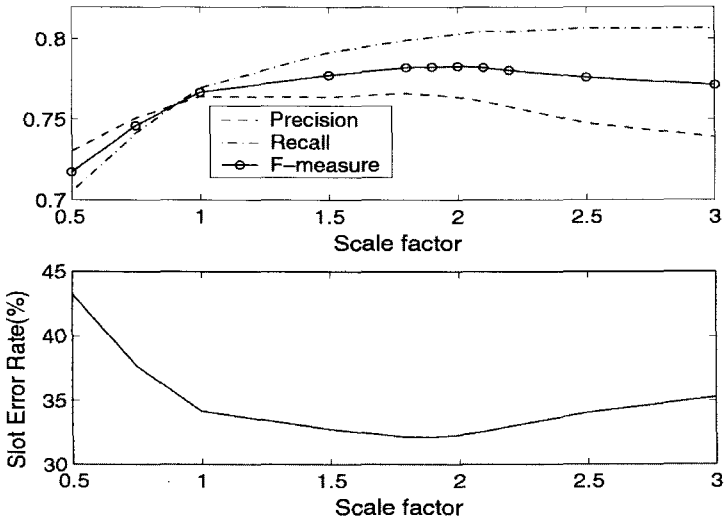
실제로 이 두 모델을 결합함으로서 자동 구두점 생성의 결과는 크게 향상된다. 두 모델이 결합된 시스템(LM+CART)은 scale factor가 2.0인 경우, F-measure로 0.7830, SER로 32.30%의 결과를 보인다. 이 때 precision은 0.7683이 되고, recall은 0.8031이 된다. 기준 전사에 대한 자동 구두점 생성의 결과는 <표 3>에 정리되어 있다.

<표 3> 기준 전사에 대한 구두점 생성 결과

시스템 명	Precision	Recall	F-measure	SER(%)
LMOnly	0.5966	0.5488	0.5717	72.25
CARTOnly	0.5383	0.7417	0.6238	71.71
LM+CART(α=2.0)	0.7638	0.8031	0.7830	32.30

LM+CART의 성능은 scale factor가 변화함에 따라서 변화하게 된다. <그림 3>은 F-measure, precision, recall 그리고 SER이 scale factor가 변화함에 따라서 변화하는 것을 보여준다. Prosodic feature 모델에 대한 scale factor값이 커짐에 따라 recall값이 커지게 되는데, 이는 CARTOnly에 대해서 recall이 precision에 비해서 훨씬 크기 때문이다. F-measure는 scale factor가 2.0인 경우 최대값인 0.7830을 가진다. SER은 scale factor가 1.8인 경우 최소값인 32.12%를 가진다.

만약 scale factor라는 개념이 본 연구에서 도입되지 않았다면, 언어모델로부터의 확률과 prosodic feature 모델로부터의 확률은 1:1로 결합되었을 것이다. Scale factor값으로 1.0이 사용했을 때, F-measure는 0.7668이 되고, SER은 34.16%가 된다. Scale factor가 도입됨으로서 F-measure는 0.0162(2.11% relative), SER은 2.04%(5.97% relative) 만큼 개선되었다.



<그림 3> Scale factor의 변화에 따른 LM+CART의 구두점 생성 결과

제안한 구두점 생성 방법은 음성인식기의 1-best 결과에도 적용 가능하다. 1-best 결과가 우선 time align된다. Time alignment 정보에 따라 prosodic feature 값들이 각 단어의 끝에서 측정된다. 3장에서 기술된 구두점 생성 방법과 같이, 1-best 결과에 대해 가장 최적의 구두점 열들이 prosodic feature 모델과 구두점을 포함한 텍스트에서 학습된 언어 모델에 의하여 생성된다.

실시간의 10배 이하에서 작동하는 (under 10 times real time. 이하 10xRT) 뉴스 자료에 대한 HTK 시스템[15]을 음성인식기의 1-best 결과 생성에 이용했다. 이 시스템의 가장 첫 번째 단계는 segmentation 단계인데, 이 단계에서 연속되어지는 뉴스자료 stream이 하나의 화자가 하나의 audio type(예: wide-band, narrow-band 등)으로 이야기하는 segment들로 분리가 된다.

HTK 시스템에서 음성인식은 two pass로 진행이 되는데, 각 pass에서는 cross-word triphone을 기반으로 한 decision-tree state cluster된 HMM과 N-gram 언어모델을 사용한다. 첫 번째 pass에서는 gender-independent(그러나 bandwidth에는 specific한) HMM과 trigram 언어모델을 이용해서 각 segment에 대해서 초기 전사를 생성한다. 그 후 gender-dependent HMM을 이용해서 각 segment별로 화자의 성별이 결정된다. 그 후 각 segment 별로 unsupervised Maximum Likelihood Linear Regression(MLLR)[16] 변환을 첫 번째 pass에서 만들어진 초기 전사와 gender-dependent HMM을 이용해서 수행한다. 변환된 HMM과 4-gram 언어모델을 이용해서 두 번째 pass에서 최종 음성인식 결과를 생성한다.

뉴스자료에 대한 HTK 시스템의 구현에 대한 세부사항은 Woodland와 Hain 등의 연구[17][18]에 자세히 기술되어 있고, 10xRT HTK 시스템의 구현에 대한 세부

사항은 Odell과 Woodland의 논문[19]에 자세히 기술되어 있다. HTK 시스템의 속도를 높이기 위해서 10xRT 시스템에서는 단순화된 음향 모델과 단순화된 decoding 방법을 이용했다.

10xRT 시스템을 이용해서 음성인식이 TestBNAcoustic98에 수행되었다. 단어오류율은 16.7%로 측정되었다. 동일한 테스트 자료에 대해 NIST 1998 Hub-4 방송 뉴스 벤치마크 테스트에서 10xRT HTK 시스템의 보고된 단어오류율은 16.1%로서 [20] 본 연구에서의 음성인식 결과와 0.6%의 차이가 있다. 본 연구에서 사용한 10xRT HTK 시스템은 1998 Hub-4 방송 뉴스 벤치마크 테스트에서 사용된 10xRT 시스템과 category-based 언어모델 [21] 사용 여부, 언어 모델 학습 자료 양, 어휘 크기에서 차이가 있다.

이 10xRT 시스템이 생성한 1-best 음성인식 결과를 LM+CART_ASR1Best라 하자. LM+CART_ASR1Best의 scale factor의 변화에 따른 구두점 생성의 F-measure와 SER의 변화 추이는 기준 전사에 대한 자동 구두점 생성 시스템(LM+CART)의 구두점 생성 결과에서의 변화 추이와 유사했다. LM+CART_ASR1Best의 SER은 scale factor가 1.93에서 최소값을 가지며, F-measure는 scale factor가 2.10에서 최대값을 가진다. <표 4>는 LM+CART_ASR1Best의 결과를 보여준다.

<표 4> LM_CART_ASR1Best의 자동 구두점 생성 결과

시스템 명	WER(%)	Precision	Recall	F-measure	SER(%)
LM+CART_ASR1Best(α=2.10)	16.71	0.5329	0.4304	0.4762	88.32

LM_CART_ASR1Best의 자동 구두점 생성 결과는 <표 3>의 기준 전사에 대한 구두점 생성결과와 비교 했을 때, WER 16.71%를 고려하더라도 많은 차이가 나고 있다.

본 연구에서 구두점의 발음은 silence로 발음 사전에 등록되었다. TestBNAcoustic98에 대해서 기준 단어열을 이용해서 alignment를 수행하게 되면, 실제 구두점에서 silence의 유무를 판별할 수 있다. TestBNAcoustic98에 대해서 각 구두점 위치에서의 silence 유무를 조사해 본 결과, 마침표와 쉼표의 90%에서 silence가 있었지만, 쉼표의 약 40%에서는 pause가 없었다. 또한 구두점이 아닌 단어들 중 약 15%의 단어의 끝에서 pause가 검출되었다. 따라서, 음성인식 결과에 대한 구두점 생성의 결과를 향상시키기 위해서는 구두점의 발음에 대한 좀 더 정확한 가정이 필요하다.

6. 결 론

본 논문에서는 음성 자료에 대한 전사가 주어진 경우, 이에 대해 prosody 정보와 구두점을 포함한 확장된 언어모델을 이용하여 자동으로 구두점을 생성하는 방법을 제시했다. 이 방법을 기준 전사를 이용해서 검증해 본 결과, prosody 정보가 언어 모델에 비해서 자동 구두점 생성에 더 유용했고, 이 두 정보원이 결합될 때, 기준 전사의 자동 구두점 생성에 대해서 F-measure로 0.7830의 결과를 얻었다.

같은 방법이 음성인식기의 1-best 결과를 이용해서 검증되었다. 음성인식기의 1-best 결과가 우선 time align된다. Time alignment 결과를 바탕으로 prosodic feature들이 계산되고, 기준 전사에 대해서 적용된 자동 구두점 생성 방법과 동일한 방법으로 prosodic feature 모델과 언어모델을 이용하여 최적의 구두점 열이 생성된다. Prosodic feature 모델에 대한 가중치의 변화에 따른 F-measure 및 SER의 변화 추이는 기준 전사에 대한 실험에서의 변화 추이와 유사하게 나타났다.

참 고 문 헌

[1] D. Beeferman, A. Berger, J. Lafferty, "Cyberpunc: A lightweight punctuation annotation system for speech", *Proc. ICASSP*, pp. 689-692, 1998.

[2] C. Chen, "Speech recognition with automatic punctuation", *Proc. European Conference on Speech Communication and Technology*, pp. 447-450, 1999.

[3] Y. Gotoh, S. Renals, "Sentence boundary detection in broadcast speech transcripts", *Proc. International Workshop on Automatic Speech Recognition*, pp. 228-235, 2000.

[4] D. Palmer, M. Hearst. "Adaptive multilingual sentence boundary disambiguation", *Computational Linguistics*, Vol. 23, No. 2, pp. 241-269, 1997.

[5] E. Shriberg, R. Bates, et. al., "Can prosody aid the automatic classification of dialog acts in conversational speech?", *Language and Speech*, Vol. 41, Nos. 3-4, pp. 439-487, 1998.

[6] M. Fach, "A comparison between syntactic and prosodic phrasing", *Proc. European Conference on Speech Communication and Technology*, pp. 527-530, 1999.

[7] S. Abney, "Chunks and dependencies: Bringing processing evidence to bear on syntax", *Computational Linguistics and the Foundations of Linguistic Theory*, pp. 145-164, 1995.

[8] K. Silverman, M. Beckman, et. al., "ToBI: A standard for labelling English prosody", *Proc. ICSLP*, pp. 867-870, 1992.

[9] P. Taylor, S. King, et. al., "Intonation and dialog context as constraints for speech recognition", *Language and Speech*, Vol. 41, Nos. 3-4, pp. 489-508, 1998.

[10] L. Breiman, J. Friedman, et. al., *"Classification and Regression Trees"*, Wadsworth and Brooks, 1983.

[11] L. Rabiner, B. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[12] J. Makhoul, F. Kubala, et. al., "Performance measures for information extraction", *Proc. DARPA Broadcast News Workshop*, pp. 249-252, 1999.

[13] A. Stolcke, E. Shriberg, et. al., "Combining words and speech prosody for automatic topic segmentation", *Proc. DARPA Broadcast News Workshop*, pp. 61-64, 1999.

[14] D. Hakkani-Tur, G. Tur, et. al., "Combining words and prosody for information extraction from speech", *Proc. European Conference on Speech Communication and Technology*, pp. 1991-1994, 1999.

[15] P. Woodland, "The development of the HTK broadcast news transcription system: An overview", *Speech Communication*, Vol. 37, Nos. 1-2, pp. 47-67, 2002.

[16] C. Leggetter, P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, Vol. 9, pp. 171-185, 1995.

[17] P. Woodland, T. Hain, et. al., "The 1997 HTK broadcast news transcription system", *Proc. Broadcast News Transcription and Understanding Workshop*, 1998.

[18] P. Woodland, T. Hain, et. al., "The 1998 HTK broadcast news transcription system: Development and results", *Proc. DARPA Broadcast News Workshop*, pp. 265-270, 1999.

[19] J. Odell, P. Woodland, T. Hain, "The CUHTK-Entropic 10xRT broadcast news transcription system", *Proc. DARPA Broadcast News Workshop*, pp. 271-275, 1999.

[20] D. Pallett, J. Fiscus, et. al., "1998 broadcast news benchmark test results: English and non-English word error rate performance measures", *Proc. DARPA Broadcast News Workshop*, pp. 5-12, 1999.

[21] T. Niesler, E. Whittaker, P. Woodland, "Comparison of part-of-speech and automatically derived category-based language models for speech recognition", *Proc. ICASSP*, pp. 177-180, 1998.

접수일자: 2007년 2월 16일
게재결정: 2007년 3월 24일

▶ 김지환(Ji-Hwan Kim)
주소: 137-724 서울시 서초구 우면동 16번지
소속: LG전자기술원 정보기술연구소 IST 그룹
전화: 02) 526-4164
E-mail: kimjihwan@lge.com