2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

MAP 확률을 이용한 끊어읽기 강도 예측

김상훈, 박준, 이영직
한국전자통신연구원, 음성언어팀

Break Strength Prediction Using Maximum a Posterior Probability

Sanghun Kim, Jun Park, and Youngjik Lee
Spoken language Processing Team, Electronics and Telecommunications Research Institute
{ksh, junpark, ylee}@etri.re.kr

본 논문은 자연스러운 합성음 생성을 위한 끊어읽기 강도 예측에 관한 것으로, 문장에 대한 품사열이 주어졌을 때 posteriori 확률을 최대화 하는 끊어읽기 강도를 비터비 디코딩으로 예측한다. 훈련용 데이터는 여성화자 1인이 발성한 2,100 문장이며, 음성데이타로부터 휴지길이(pause)에 따라 끊어읽기 강도를 2단계로 할당하고, 텍스트에서는 30개의 품사 태그심볼을 이용하여 형태소분석 및 태깅을 수행하였다. 관측확률은 3개 연속하는 품사열이 발생할 확률로 하고 끊어읽기 강도 천이확률은 bigram으로 했을 때, cross validation 방법으로 성능 평가를 수행하였다. 평가결과, 훈련데이타에 대해서는 89.7%, 테스트 데이터에 대해서는 84.9%의 예측정확률을 보였다.

1. 연구목적

ETRI 합성기는 대량의 음성 코퍼스로부터 운율이 살린 합성단위를 최적 선택하여 연결하는 합성방식을 채택하고 있다. 그림 1은 합성데이타 제작에서부터 합성단위 선정까지를

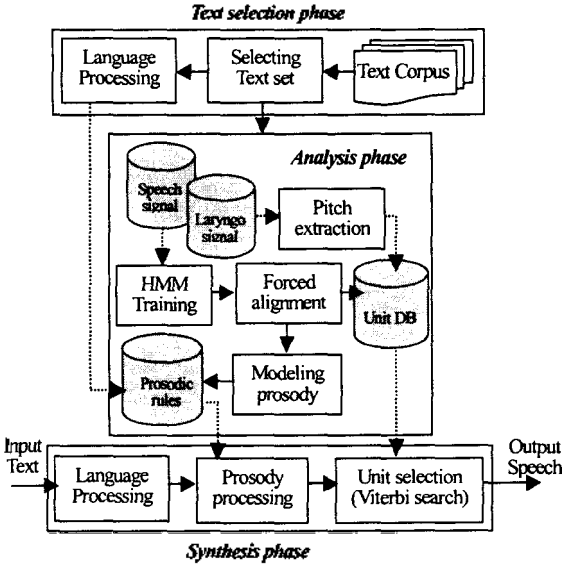

그림 1: 학습형 음성합성기 구조

도식화 한 것이다. 이 합성기에서는 억양, 지속시간, 에너지 등 운율처리를 하고 있지 않으며, 다만 어절 경계 유형, 즉 끊어읽기 강도를 예측하는 기능만 가지고 있다. 이는 인간의 지각에 크게 영향을 미치는 주요 운율변화가 주로 어절 경계에서 일어나기 때문에 어절 경계에서의 적절한 합성단위 선정만으로도 합성음의 자연성을 살릴 수 있다. 따라서 이 합성기에서 사용하고 있는 합성단위는 어절 경계의 휴지길이에 따라 합성단위가 세분화되어 있어 끊어읽기 강도에 따른 운율현상이 합성단위에 반영되어 있다. 합성할 때에는 어절 경계에서 사용할 합성단위와 어절간 할당해야 할 휴지길이를 선정하기 위해 텍스트의 품사정보를 이용하여 통계적인 방법으로 경계유형을 예측한다. 초기에는 품사 bigram과 trigram을 이용하여 예측하였으나 여전히 합성음이 불안정하였고, 적절한 끊어읽기가 되지 못한바 이 논문에서는 더 강인한 끊어읽기 예측을 수행하고자 한다.

2. 실험조건 및 데이터베이스

경계유형은 휴지길이만으로 결정되는 것은 아니다. 휴지정보 외 억양, 지속시간 정보가 경계유형을 결정짓는데 중요한 특징이다. 그러나 휴지정보는 억양, 지속시간 정보에 비해 경계유형을 찾는데 용이하게 추출할 수 있는 특징으로서 또한 경계유형 결정에 사실상 절대적으로 기여하기 때문에 이 실험에서는 휴지길이 정보만 이용하여 훈련용 데이터에 끊어읽기 정보를 할당하였다. 끊어읽기 강도는 그림 2와 같이 휴지길이의 분포에 따라 4단계(1단계: 휴지가 없음, 2단계: 0<pause<40msec, 3단계: 40msec<pause<220msec, 4단계: pause>220msec)로 나누었으나 이 논문에서는 실험을 단순화하기 위해 휴지가 있고("break"), 없고("non break") 에 따라 2단계로 구별하였다.

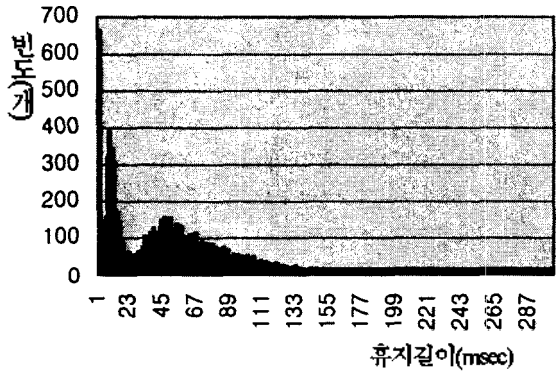

그림 2: 어절간 발생하는 휴지길이 분포

ToBI에서의 break index와 비교하자면, 이 실험에서 사용한 "break"는 break index 4에 해당하고, "no break"는 break index 0,1,2,3에 해당한다. break index 2인 경우, 휴지가 있을 경우 "break"로 포함한다. 이 실험에서는 break index 3(intermediate phrase or accentual phrase)는 따로 단계를 구별하여 사용하지 않았다. 훈련용 데이터는 여성화자 1인이 발성한 2,100여 문장이고 자동 음소레이블링으로 음소구간 및 휴지길이를 추출하였다. 휴지길이는 다시 레링고 데이터를 이용하여 보정하고 최종 수동으로 수정되었다.

텍스트는 형태소 해석기를 이용해 자동 태깅을 수행하고 수동 보정하였다. 당초 사용된 태거 세트(품사정보)는 약 60개로 훈련데이타량에 비해 세분화되어 있음으로 이를 30개로 줄여 사용하였다. 30개로 분류된 태거세트는 표 1과 같다.

표 1: 끊어읽기용 태거 세트

분류	태그명	분류	태그명
기호	s	단어접속부사	mac
외국어	f	독립언	I
서술성명사	ncp	주격조사	jcs
비서술성명사	ncn	관형격조사	jcm
고유명사	nq	호격조사	jcv
단위성의존명사	nbu	접속격조사	jcj
비단위성의존명	nbn	보조사	jx
대명사	np	선어말어미	ep
숫자	nnn	연결어미	ec
일반수사	nng	관형사형전성어미	etm
동사	pv	명사형전성어미	etn
형용사	pa	종결어미	ef
일반 관형사	mmg	명사파생접미사	xsn
수관형사	mmc	용언파생접미사	xsd
일반부사	mag	부사파생접미사	xsa

3. 끊어읽기 예측 알고리즘

- HMM-like model

HMM에서 정의되는 state와 각종 파라미터(관측확률, 천이확률)는 끊어읽기 강도 예측 알고리즘에서 다음과 같이 매핑될 수 있다.

State=끊어읽기 강도(Break strength)
Transition probability= 끊어읽기 강도의 천이 확률
Observation probability=품사(POS)발생 확률
Recognition= Viterbi decoding,

- Probability model

품사열(W)이 주어졌을 때, 최적 끊어읽기 강도(B)가 할당될 MAP(Maximum a Posteriori) 확률은

$$\hat{Q(B)} = \arg\max_{B_1^n} \{P(B_1^n \mid W_1^n)\}$$

이고, Markov chain rule에 따라 풀어쓰면,

$P(B_1^n \mid W_1^n) = P(B_1 \mid W_1^n) \times P(B_2 \mid B_1, W_1^n) \times P(B_3 \mid B_1, B_2, W_1^n), \cdots, P(B_n \mid B_1^{n-1}, W_1^n)$ 수 있다.

이 된다. 여기서 Bayes rule을 적용하면,

$$P(B_1^n \mid W_1^n) = \prod_{i=1}^{n} P(B_i \mid B_0^{i-1}, W_1^n) = \prod_{i=1}^{n} P(B_i \mid B_0^{i-1} \mid W_1^n)$$

이며, $P((B_i \mid B_0^{i-1}) \mid W_1^n)$ 은 다시

$$P((B_i \mid B_0^{i-1}) \mid W_1^n) = \frac{P(W_1^n \mid (B_i \mid B_0^{i-1}))P(B_i \mid B_0^{i-1})}{P(W_1^n)}$$

로 쓸수 있고, $P(W_1^n \mid (B_i \mid B_0^{i-1}))$ 는 B_0^{i-1} 에 독립이라고 가정하고(Conditionally independent of the past), 관측확률(Observation probability) $P(W_1^n \mid B_i)$ 을 품사열 W_{i-2} W_{i-1} W_i 로 근사화하면,

$$P(W_1^n \mid (B_i \mid B_0^{i-1})) \approx P(W_1^n \mid B_i) \approx P(W_{i-2}, W_{i-1}, W_i \mid B_i)$$

로 쓸수 있다.

n-gram 천이확률 $P(B_i \mid B_0^{i-1})$ 를 1차 Markov model로 가정하면

$$P(B_i \mid B_1^{i-1}) \approx \prod_{i=2}^{n} P(B_i \mid B_{i-1})$$

로 근사화 되고, 최종 Criterion $Q(\bar{B})$을 다시 정리하면,

$$Q(\bar{B}) = \arg\max_{\bar{B}} P(B_1^n \mid W_1^n)$$

$$\approx \arg\max_{\bar{B}} P(W_1 \mid B_1) \prod_{i=2}^{n} P(W_{i-2}, W_{i-1}, W_i \mid B_i) P(B_i \mid B_{i-1})$$

로 근사화할 수 있다.

● Viterbi Decoding

관측 품사열 $W=W_1W_2W_3\ldots W_T$ 가 주어졌을 때, 최적 state 끊어읽기 강도열 $B=\{B_1B_2\ldots B_T\}$은 Viterbi decoding에 의해 구할

$$\delta_t(i) = \max_{B} P(B_1B_2 \cdots B_t = i, W_1W_2 \cdots W_t) \mid \lambda)$$

$$\delta_{t+1}(i) = [\max_i \delta_t(i)a_{ij})]b_j(W_{t+1})$$

i) 모든 state i(1<i<N)에 대해 초기화

$$\delta_1(i) = \pi\delta_i b_j(W_1)$$

$$\psi_1(i) = 0$$

ii) 각 state에서 최대 누적 확률값과 이전 state의 정보를 저장

$$\delta_t(i) = \max_i[\delta_{t-1}(i)a_{ij})]b_j(W_t)$$

$$\psi_t(i) = \arg\max_i[\delta_{t-1}(i)a_{ij})]$$

iii) 최종 state에서 최대 누적확률값을 나타내는 state $\hat{B}_n$ 으로부터 역경로(backward path)로 시간 k에서 최대 누적 확률값을 나타내는 state로 traceback

$$\hat{B}_n = \max_{k \vee A} \delta_n(k)$$

$$\hat{B}_{n-1} = \psi_i(\hat{B}_n)$$

● 3개 연속 품사정보를 이용한 관측확률

예측성능을 높이기 위해 3개 연속 품사열을 이용한다. 또한 합성에서는 이미 문장에 대한 품사열을 알고 있기 때문에 현재 state(i)에서 과거 i-1, i-2의 품사열과 미래 i+1에 대한 품사열 정보를 이용한다. 이는 4개 연속 품사정보열을 근사화하는 역할을 한다.

Observation _ probability

$$\approx P(W_{i-2}, W_{i-1}, W_i \mid B_{i-1}) \times P(W_{i-1}, W_i, W_{i+1} \mid B_{i-1})$$

4. 실험결과

여성화자 1인이 발성한 2,100 여 낭독체 문장으로부터 훈련, crossvalidation test한 결과는 표 2와 같다.

표 2: 3 consecutive POS observation probability+bigram transition probability 예측 성능

	Juncture Correct	Break Correct	Juncture Insertion
Closed test	89.7%	79.1%	5.5%
Open test	84.9%	70.9%	8.0%

표 2에서 Juncture Correct는 전체 어절경계(break+non break) 중 올바르게 예측된 어절경계의 정도를 의미하고, Break Correct는 "non break"을 제외한 전체 "break" 중 올바르게 예측된 "break"의 정도를 의미한다. Juncture Insertion은 전체 "non break"중 "break"가 삽입된(잘못 예측된) "non break'의 정도를 말한다.

5. 결론

이 논문에서는 MAP 확률을 최대화하는 끊어읽기 강도예측 알고리즘을 "break", "non break"의 2단계 끊어읽기 강도 예측실험에 적용하였다. 이 실험을 단순하게 하기 위해 2단계만 사용했으나 실제 합성기에서는 이보다 더 세밀한, 적어도 4단계 이상은 끊어읽기 강도 또는 어절 경계 패턴이 되어야 하고 이를 텍스트에서 추정할 수 있어야 한다. 그리고 끊어읽기 혹은 경계유형을 정할 때 ToBI 시스템의 break index에 기반한(clitic group: 0, default: 1, mismatch: 2, accentual phrase: 3, intonational phrase: 4) 유형 분류가 되어야 할 것이다. 또한 끊어읽기 패턴이 사람마다 다르고 꼭 한 가지 패턴만이 있는 것은 아니기 때문에(반드시 끊어야 할 경우와 끊어질수도 있고 아닐 수도 있는 경우가 있음) 다수 화자의 끊어읽기 패턴에 대한 연구도 필요하며, 지금과 같이 한가지 패턴과 비교하여 결과를 내는 것보다 다수 화자 패턴이 성능 평가에 반영해야 할 것이다.

참고문헌

[1] Beckman, M.E., Ayers, GM., 1994. Guidelines to ToBI labelling. Version 2.0.

[2] Beckman, M.E., Hirschberg, J.,1994. The ToBI Annotation Conventions.

[3] Black, A.W., Campbell, N., 1995. Optimizing Selection of Unit from speech database concatenative synthesis. In: Proceedings of EUROSPEECH'95, Madrid, vol. 1 pp.581-584.

[4] Hauptmann, A.G, 1993. SPEAKEZ: A First Experiment in Concatenation Synthesis from a Large Corpus. In: Proceedings of EUROSPEECH'93, Berlin, pp.1701-1704.

[5] Hirose, K., 1998. Processing of Prosodic Information. In: Journal of Signal Processing, vol.2, No. 6. Pp.415-423

[6] Hirose, K., Fujisaki, H., 1993. A system for the synthesis of high quality speech from texts on general weather conditions. In: IEICE Trans. On Fundamentals, vol. E76-A, No.11, pp.1971-1980.

[7] Hunt, A.J., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of ICASSP'96, Atlanta, pp.373-376.

[8] Kim, S.H., et al, 1999. An experiment for improving stability of sound and downsizing synthesis database. Proceedings of ICSP'99, Seoul, vol 1, pp. 209-212.

[9] Kim, S.H., Lee, H.S., and Kim, H.R., 1996. An Effectiveness of Automatic Labeling using Speech Recognizer. In: Proceedings of SICOPS96,seoul, SESSON 3.6.

[10] Sanders, E., Taylor, P., 1995. Using Statistical Models to Predict Phrase Boundaries for Speech Synthesis. In: Proceedings of EUROSPEECH'95, Madrid, pp.1811-1814.

[11] Sagisaka, Y., 1998. Corpus based Speech Synthesis. In: Journal of Signal Processing, vol.2, No. 6., pp. 407-414.