Large scale word recognizer를 위한 음성 database - POW

임연자, 이영직
한국전자통신연구소 음성언어연구실
305-600 대전직할시 유성구 유성우체국 사서함 106호
yjlim@zenith.etri.re.kr

The Speech Database for Large Scale Word Recgnizer

Yeonja Lim, Youngjik Lee
Spoken Language Processing Section, ETRI P.O. Box 106, Yusung, Taejon, 305-600, Korea

요약

본 논문은 POW (phonetically optimized word) algorithm과 알고리즘을 통해 수행된 결과인 large scale word recognizer를 위한 POW set에 대하여 설명 하겠다.

Large scale word recognizer를 위한 speech database를 구축하기 위해서는 모든 가능한 phonological phenomenon이 POW set에 포함 되어야 한다. 또한 POW set의 음운현상들의 분포는 추출하고자 하는 모집단의 음운현상들의 분포와 유사해야 한다. 위와 같은 목적으로 다음과 같이 3가지 성질을 갖는 POW set을 추출하기위한 새로운 algorithm을 제안한다.

1. 모집단에서 발생하는 모든 음운현상을 포함해야 한다.
2. 최소한의 단어 집합으로 구성되어야 한다.
3. POW set과 모집단의 음운현상의 분포가 유사해야 한다.

우리는 약 300만 어절의 한국어 text corpus로부터 5천 단어의 고빈도 어절을 추출하고 이로부터 한국어 POW set을 추출하였다.

1 서론

Speech database는 크게 두 가지 목적에서 구성 될 수 있다. 첫째는 음성 인식률을 높이기 위하여 specific한 domain에서 database를 구축할 수 있다. 이러한 경우에는 speech database의 단어 집합은 domain에 한정 될 수 밖에 없다.

두번째는 large scale recognition system 그리고 COC(context oriented clusters)를 이용하는 음성합성 system 구축을 위한 것과 같이 general한 목적으로 speech database를 구성 할 수 있다. 두번째 경우의 목적으로 speech database를 구축하기 위해서는 특별히 transctription된 text의 집합 구성이 중요하다.

최근까지 두번째의 경우와 같은 database 구축을 목적으로 PBW algorithm이 개발되었다. PBW algorithm은 entropy의 최대화에 기본을 두어 PBW set에 발생하는 음운현상들이 equal probability를 갖게됨을 의미한다. 그러나 PBW algorithm에 의해 추출되는 PBW set의 triphone은 그 기본개념과는 거리가 있음을 PBW set의 분석을 통하여 밝혀 내었다. 위와 같은 문제를 해결하기 위하여 information theory를 기초로하는 새로운 POW algorithm을 제안하였다.

2 PBW algorithm의 문제점

General한 목적으로 구축되어진 PBW set은 Large scale word recognizer의 인식률을 높이지 못한다. 그 이유는 모집단에 존재하는 음운현상 - 앞으로는 triphone으로 하겠다. - 과 PBW set에 존재하는 triphone의 분포가 서로 다르기 때문이다. Equal probability를 갖는다는 의미는 자주 사용되는 단어의 인식률과 드물게 사용되는 단어의 인식률이 같아짐을 의미하기 때문이다. - 본 이론을 뒷받침할 논문을 공식적으로 발표할 예정임 - 따라서 음성인식률을 저하시키는 결과를 초래하게 된다. POW algorithm은 이러한 문제를 개선 하였다.

3 POW (phnetically optimized words) algorithm

3.1 POW set의 정의

POW set은 다음과 같이 정의할 수 있다. POW set은 모집단에서 발생 가능한 모든 triphone을 포함해야 한다. POW set에 있는 triphone의 분포가 모집단의 triphone과 유사해야 한다. 그리고 POW set은 최소한의 단어를 포함해야 한다.

이는 이미 발표된 본인의 논문[4]에서 서술했던 PBW algoirhtm

과 비교해 볼때 entropy 적용방법에서 차이가 있다.

POW algorihtm은 다음과 같은 순서로 세가지 조건을 모두 만족해야 한다.

1. POW set은 모집단에서 발생 가능한 모든 triphone을 포함해야 한다.
2. POW set은 최소한의 단어 집합으로 구성되어야 한다.
3. POW set에 포함된 triphone의 분포가 모집단에 존재하는 triphone의 분포와 최대한 유사해야 한다.

여기에서의 모집단은 약 300만 어절의 ETRI corpus에서 고빈도 5천 어절을 추출하여 구성 하였다.

POW algorithm은 다음과 같은 두 기본 개념이 포함 되어야 한다.

1. Modified entropy의 최대화
2. 단어 수의 최소화

3.2 Modified Entropy

POW algorithm에서 사용된 modified entropy는 다음과 같이 유도될 수 있다.

1. 모집단의 triphone을 분석하여 i번째 triphone의 빈도수 A_i, $i = 1, 2, \ldots, N$,와 그중에 최고의 빈도를 갖는 triphone의 빈도를 찾아 $\mathcal{M}_{max}$라 놓는다.
2. $B_i = \mathcal{M}_{max} - A_i$.
3. a_i는 POW를 구성하는 과정에서의 i번째 triphone의 빈도라 한다.
4. $$T = \sum_{i=1}^{N} a_i + B_i \qquad (1)$$
5. 따라서 modified entropy 식은 다음과 같다.
$$H = \sum_{i=1}^{N} (a_i + B_i)/T \log (a_i + B_i)/T \qquad (2)$$

3.3 ADD/DELETE를 이용하는 POW algorihtm

Modified entropy의 식 H를 최대로하는 POW algorithm은 다음과 같다.

- Initial set의 구성 과정
  1. 모집단을 분석하여 모집단에서 오직 한번만 출현하는 triphone을 포함하는 단어들로 initial set을 구성한다.
  2. 수작업을 통하여 initial set에서 매우 드물게 사용되거나 외래어인 경우에는 initial set로부터 단어를 삭제한다.
  3. Initial set으로 구성된 단어들을 모두 POW set에 첨가한다.
- ADD 과정
  1. Initial set 구성과정을 거친 나머지 단어들은 단어 자체의 modified entropy와 단어의 길이가 같고 서로다른 triphone으로 구성된 단어 순서로 candidate list를 구성한다.
  2. Candidate list중 POW set의 modified entropy를 최대로 하는 단어를 선택하여 POW set에 첨가한다.
  3. 1, 2의 과정을 candidate list의 단어가 POW set에 더이상의 새로운 triphone을 첨가하지 못할때까지 계속 수행한다.
- DELETE 과정
  1. ADD 과정을 거친 POW set에서 modified entropy를 감소시키는 즉 해당 단어를 삭제해도 triphone의 수를 감소시키지 않으면서 modified entropy를 최대로하는 단어를 POW set으로부터 삭제한다.
  2. 1번 과정을 modified entropy가 최대가 될때까지 수행한다.

4 비교

이미 언급했듯이 300만 어절의 ETRI corpus에서 5000 개의 고빈도 어절을 추출하여 transcription하였으며 이로부터 PBW와 POW set을 구축 하였다. POW set은 3,848개의 어절이 PBW set은 4,028개의 어절이 포함 되어있다. POW와 PBW set에 total triphone 수는 각각 25,795 개와 56,784 개가 포함 되어있다.

모집단과 각각 집단의 triphone 분포의 유사도를 측정하기 위하여 divergence를 측정 하였다. 모집단과 POW는 1.28, 모집단과 PBW는 5.62가 측정되었다. 아래의 그림1, 그림2, 그림3은 각각의 triphone분포를 histogram으로 보여주고 있다.

5 그 밖의 문제

양질의 POW set을 구성하는 문제는 잘 고안된 POW algorithm이 무엇보다 중요하다. 뿐만 아니라 그 이외에도 중요한 여러 문제가 있다.

첫째, 모집단을 추출하는 corpus의 형태이다. 이미 domain이 없는 database 구축을 목적으로 하므로 어느 장르에도 치우치지 않는 balanced 형태를 취하고 있어야 한다.

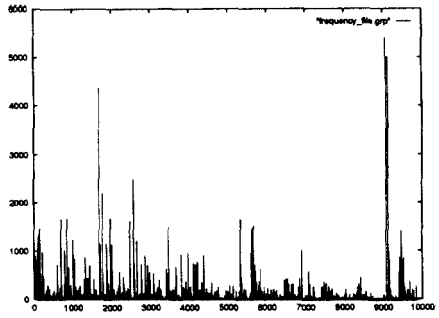

그림 1: Frequency of triphones in the population

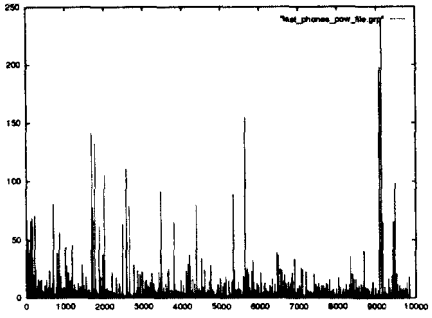

그림 2: Frequency of triphones in the POW set

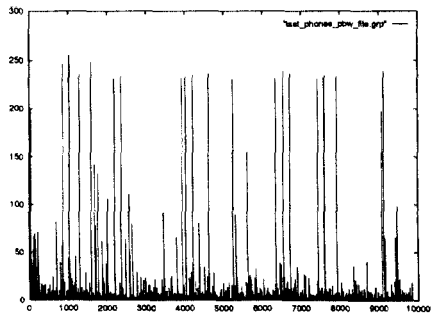

그림 3: Frequency of triphones in the PBW set

둘째, 구성된 모집단을 규칙 등을 적용하여 발음나는대로 표기하는 transcription의 문제이다. 실제로 정의된 발음규칙대로 발성하지 않는 경우가 많기 때문에 특별히 음운규칙을 보다 세분화하여 text를 transcription하는 것이 무엇보다 중요하다.
세째, 세번째는 본 연구실에서 앞으로 수행해야할 일로 단순한 transcription뿐만 아니라 보다많은 음성학적지식, 언어학적지식을 포함하는 POW set을 구축하는 것이다.

6 앞으로 할 일

POW set은 Large scale word recognizer와 COC를 이용하는 T-t-S system을 위하여 구축되었다. 따라서 본 연구실에서는 POW set을 이용하여 large scale word recognizer를 구축할 계획이 있다. 또한 보다 세밀한 음운현상과 다양한 지식을 표현해 주는 transcription tool을 구축하여 보다 정확하고 풍부한 내용을 포함하는 POS set을 올해 구축할 계획이다.

감사의 글

본 연구와 관련하여 한국통신의 부분한 지원에 감사를 드립니다.

참고 문헌

[1] K. Shikano, "Phonetically balanced word list based on information entropy," *Proceedings of Acoustical society of Japan*, 1984.

[2] Hayamizu, Tanaka, Yokayama, and Ohta, "Generation of VCV/CVC balanced word sets for speech data base," *Bulletin of Electrotechnical Laboratory*, 1986.

[3] J. Lin, "Divergence measure based on the Shannon's entropy ," *IEEE Trans. Inform. Theory*, vol. 37, pp. 145-151, Jan. 1991.

[4] Y.J. Lim, "Implementation of the POW algorithm for speech database ," *Proceedings of ICASSP-95*, vol. 1, pp. 89-92, May. 1995.