잡음환경하의 연속 음성인식을 위한
유사음소단위 분석

신광호, 임수호, 서준배, 김주곤, 정호열, 정현열
영남대학교 정보통신공학과

An Analysis on Phone-Like Units for Korean Continuous Speech Recognition in Noisy Environments

Guang-Hu Shen, Soo-Ho Lim, Jun-Bae Seo, Joo-Gon Kim, Ho-Youl Jung, Hyun-Yeol Chung
Dept. of Information and Communication Eng., Yeungnam University
E-mail: guanghosin@ymail.ac.kr

요 약
본 논문은 잡음환경 하에서의 효율적인 문맥의존 음향모델 구성에 대한 기초연구로서 잡음환경 하에서의 유사음소단위 수에 따른 연속 음성인식 성능을 비교, 평가한 결과에 대한 보고이다. 기존의 연구들[1,2]로부터 연속음성 인식의 경우 문맥존속모델은 변이음속을 고려한 39유사음소를 이용한 경우가 48유사음소를 이용하는 것보다 더 좋은 인식성능을 나타내었다. 이 연구 결과를 바탕으로 본 연구에서는 잡음환경에서도 효율적인 문맥의존 음향모델을 구성하기 위한 기초 연구를 수행하였다. 다양한 잡음환경을 고려하기 위해 White, Pink, LAB 잡음으로 실험원이 주위에 주변 사운드 수위에 따라 연속음성인식 실험을 수행하였다. 그 결과, 39유사음소를 이용한 경우가 48유사음소를 이용한 경우보다 clear 환경인 경우에 이 7%와 17% 이상한 단어인식률과 문장인식률을 얻을 수 있었으며, 각 잡음환경에서의 39유사음소를 이용한 경우가 48유사음소를 이용한 경우보다 평균적으로 17%와 28% 이상한 단어인식률과 문장인식률을 얻을 수 있어 39유사음소 단위가 한국어 연속음성인식에 더 적합하고 잡음환경에서도 유용함을 확인할 수 있었다.

1. 서론

실제 연구 환경은 매우 다양한 형태로 나타난 잡음 환경의 영향을 받게 된다. 그러므로 연속음성인식에서는 최적의 인식단위 선정뿐만 아니라 잡음환경을 고려한 음
성 데이터베이스를 사용할 필요가 있다[9]. 인식의 기본단위로서 기존의 틀록형모델에서 사용된 음소간의 변이 정보를 포함한 48유사음소단위와 변이정보를 제외시켜 음소단위에 가깝게 재정의 39유사음소단위를 기준으로 각각 틀록의 음직 모델을 작성하여 최적의 인식단위를 고려할 필요가 있다[5]. 따라서 학습 데이터에 점점 온반응을 고려하여, 3가지 점학(White, Pink, LAB)을 신호 대 검음비(Signal to Noise Ratio) 5dB, 10dB, 15dB 레벨로 응용에 부가하여 연속응용 데이터베이스를 사용하되록 했다. 이러한 다양한 환경을 고려한 데이터베이스를 사용하여 틀록의 음향모델 작성방법인 HM-Net으로 연속응용인식에 적합한 음소 수에 대해 검토하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서 연속응음인식을 위한 기본 인식단위를 정의하고 3장에서 실험에 사용된 장음이 부가된 음성데이터베이스에 대해서 설명한다. 4장에서는 인식실험을 통해 유사음소 단위별 인식결과를 검토하고, 마지막으로 5장에서 결론을 맺도록 한다.

2. 연속응음인식을 위한 기본 인식단위

유사음소 단위는 최소 인식단위로 많이 사용되며 기존적인 음소에 변이음을 포함하고 있는 음소이다. 음향학적 및 음성학적 유사성이 큰 경우에는 음소와 유사음소 단위는 동일하게 취급될 수 있지만 그렇지 않을 경우에는 큰 차이가 있다. 48 유사음소는 틀록형 음향 모델을 작성할 때 기본 음소만으로는 부족한 음성적인 변이음을 추가하여 정의할 것이다[8]. 하지만 틀록의 음향 모델인 HM-Net 음향 모델은 음속 데이터에 나타나는 수많은 변형 및 후행음소가 결합되어 다양한 종류의 변이음 모델이 자동 생성되기 때문에 기본 음소 단위에 변이음을 추가할 필요성이 없게 된다. 그러므로 틀록의 음향 모델을 작성하기 위한 변형 및 후행음소의 중심 음소가 되는 기본 유속음소에는 틀록형인 경우에서 유효한 변이음을 고려할 필요가 없게 된다. 불필요한 기본 유속음소의 증가는 틀록의 음향 모델 작성에서 부족한 음성 데이터의 훈련 효과를 분산시켜 모델의 강건성을 저하시키는 원인이 된다.

39유사음소 단위는 변이음을 포함하지 않는 음소청의 각기 다른 음소음소 단위이다. 연속응음인식에서는 보다 높은 인식능력을 위해서 음소의 좌우 정동 정보를 사용하기 때문에 이음 음성의 변이정보를 모두 포함하게 된다. 이러한 점을 고려하여 변이음을 정정한 음소를 제외 시킨 것이 39 유사음소 단위이다. 표 1은 기존의 48유사음 소에 대해 나타낸다. 48유사음소의 (d, g, b, z, hh, l) 계열은 표 2에서와 같은 경우로 취급하여 총 39유사음소로 재정의했다. 39유사음소 단위는 음성데이터의 부족한 학습데이터의 훈련효과를 분산시키는 것을 줄일 수 있다. 그러므로 제한된 학습 데이터에서 더 많은 학습 데이터를 확보하게 되어 좀 더 강건한 모델을 학습하게 된다[1, 2].

<table>
<thead>
<tr>
<th>표 1. 48유사음소단위</th>
</tr>
</thead>
<tbody>
<tr>
<td>구분</td>
</tr>
<tr>
<td>모음</td>
</tr>
<tr>
<td>자음</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>표 2. 48유사음소와 39유사음소의 비교</th>
</tr>
</thead>
<tbody>
<tr>
<td>구분</td>
</tr>
<tr>
<td>비고</td>
</tr>
<tr>
<td>g, d, b, z, hh</td>
</tr>
<tr>
<td>g*, d*, b*, z*, hh*</td>
</tr>
<tr>
<td>l</td>
</tr>
</tbody>
</table>

3. 장음환경에서의 음성 표현

장음은 보통 백색음소(White noise)와 유색음소(Colored noise)로 구분된다. White 장음은 스펙트럼이 모든 주파수영역에 균일하며, 시간영역에서 샘플값이 서로 상관성이 없다. 이와 달리 Pink 장음은 주파수영역의 스펙트럼이 일정하지 않은 유색임으로, 다양한 분야에서 활용할 수 있다. LAB법음은 연구실환경에서 녹음된 배경음이며, 이 장음은 기계적인 장음, 사람의 주위 배경음에서의 대화소리 등 여러 가지 배경음이 포함되어 있다. 배경음에서 기계적인 원인에 의해 발생되는 저주파
작용들은 전력 스펙트럼을 감소시키는 경향이 있으며, 주변 환경 작용에 의해 발생되는 스펙트럼은 높은 주파수 특성을 가지며, 일반적으로 불규칙한 특성을 가진다.

평균이 영역 신호를 가정할 때, 신호 대 잡음비는 식 (1)과 같이 정의될 수 있다.

$$SNR = 10 \log \frac{E_s}{E_n}$$  \hspace{1cm} (1)

여기서 $E_s$ 와 $E_n$ 은 음성신호와 잡음의 평균 에너지이다. 입력 음성신호를 $s(t)$, 선형 시변변 필터를 $h(t)$, 부가잡음은 $n(t)$로 하면, 일회차 음성신호 $x(t)$는 시간영역에서 식 (2)와 같이 표현할 수 있다.

$$x(t) = s(t) * h(t) + n(t)$$  \hspace{1cm} (2)

본 논문에서는 다양한 잡음환경을 고려하기 위해 White, Pink, LAB 잡음은 신호 대 잡음비 5dB, 10dB, 15dB 레벨로 음성에 부가한 후 각 유사성소단위 수에 따른 연속음성인식 실험을 수행하였다.

4. 인식 실험 및 결과

본 논문에서 사용한 음성 데이터는 KAIST무역 상담용 DB이다. 잡음환경을 고려하기 위해 White, Pink, LAB잡음 을 깜빡한 음성 데이터에 부가하였으며, 발생환자 중 100명 분에서 90명 분을 학습데이터로 이용하였고, 나머지 10명 분으로 화자독립 인식실험을 수행하여 유효성을 비교 검토하였다. 인식을 위한 음향모델은 2000개의 8초 단위의 HMM모델을 이용하였으며, 음성인식 알고리즘은 Word-pair 문법을 인식 문법으로 하는 One-Pass Viterbi 알고리즘을 사용하였다. 사용한 음성 데이터의 분석조건은 표 3과 같다.

<table>
<thead>
<tr>
<th>주파수</th>
<th>8kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>양자화</td>
<td>16bit</td>
</tr>
<tr>
<td>프레임 길이</td>
<td>25ms</td>
</tr>
<tr>
<td>프레임 주기</td>
<td>10ms</td>
</tr>
<tr>
<td>분석장</td>
<td>Hamming Windows</td>
</tr>
<tr>
<td>특징 파라미터</td>
<td>12차 LPC-MEL cepstrum*delta power + 1,2차의 회귀 계수 = 39차원</td>
</tr>
</tbody>
</table>

1) 무잡음 환경하의 연속음성인식 실험

무잡음환경의 White, Pink, LAB 잡음의 영향을 확인하며, 8kHz의 주파수에서 단어 인식률과 문장인식률을 표 4에서 나타낸다. 단어인식률의 경우 약 7%의 성능차를 보였으며, 문장인식률의 경우 약 17%의 성능차를 보였다. 이 실험 결과로부터 39음소가 연속음성 및 변이음성상이 저주 발생하는 연속음성인식 환경에서 오인율을 유발하는 문제를 48음소보다 더 효율적으로 처리할 수 있음을 알 수 있었다. 연속음성인식에서 39음소가 48음소에 비해 더 적절한 음소 체계임을 인식 성능을 통해 확인할 수 있었다.

### 표 4. 연속음성인식율

<table>
<thead>
<tr>
<th>PLU</th>
<th>39</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>단어</td>
<td>97.69</td>
<td>90.64</td>
</tr>
<tr>
<td>문장</td>
<td>88.83</td>
<td>71.56</td>
</tr>
</tbody>
</table>

2) 잡음환경하의 연속음성인식 실험

White, Pink, LAB 잡음은 신호 대 잡음비 5dB, 10dB, 15dB 레벨로 고정한 음성에 부가한 후 각 유사성소단위 수에 따른 연속음성인식 실험을 수행하였다. 인식 결과는 그림 1,2,3에서 나타내었다.
전형 결과, 잡음환경에서 39유소소를 이용한 경우가 48유소소를 이용한 경우보다 평균적으로 17%와 28% 향상된 단어인지율과 문장인지율을 얻을 수 있었다. 이는 39유소소 단위가 한국어 연속음성인식에 더 적합하고, 잡음환경의 연속문장인식에서도 더 효과적임을 확인할 수 있다.

5. 결 론

본 논문은 잡음환경에서의 한국어 연속음성인식에 효과적인 문맥의존 음향모델 수에 대한 연구로서 유사음 소단위 수에 따른 인식 성능을 비교, 평가 하였다. 연속 음성인식에 이용되는 문맥중복모델의 경우 변이음을 고려하여 모델이 작성되므로 이를 고려하면 기본 음소를 48음소로부터 39음소로 줄일 수 있다. 39음소의 인식에 대한 유 효성을 확인하기 위하여 48음소의 인식성능 비교 평가를 수행하였다. 또한, 실제 잡음환경에서도 유효한 문맥의존 음향모델을 생성하기 위해서 잡음이 부가된 음성데이터베이스를 구성하였으며, 이 음성 데이터베이스를 각 음소별 HMM-Net음향모델에 학습시켜 연속음성인식 실험을 수행하였다.

실험 결과, 잡음환경의 연속음성인식에서 단어인지율은 약 7%, 문장인지율은 약 17%의 인식성능향상을 보였다. 잡음환경의 연속음성인식에서도 평균적으로 39유소 소단위모델 기본음소로 사용하였을 경우, 단어 인식율은 약 17%, 문장인식율은 약 28%의 성능향상을 보였다. 따라서 39음소가 발음변이가 변화하여 일어나는 연속음성 인식 환경에서 48음소보다 효과적인 음소구성임을 알 수 있었으며, 잡음환경하의 연속음성인식에서도 효과적임을 확인할 수 있었다.

참고 문헌