음소인식기와 음소결합확률모델을 이용한 언어식별시스템

이대성°, 김세현, 오영환
한국과학기술원 전자전산학과 전산학전공

Language Identification System using phoneme recognizer and phonotactic language model

Dae-Seong Lee° , Se-Hyun Kim, Yung-Hwan Oh
Division of Computer Science
Department of Electrical Engineering & Computer Science
Korea Advanced Institute of Science and Technology
E-mail: {dslee, shkim, yhoh}@bulsai.kaist.ac.kr

요약

본 논문에서는 음소인식기와 음소결합확률모델을 이용하여 전화음성을 대상으로 입력음성이 어느 나라 말인지를 식별할 수 있는 언어식별시스템을 구현하였고 성능을 실험하였다. 시스템은 음소인식기로 입력음성에 대한 음소열을 인식하는 과정, 인식된 음소열을 이용하여 인식대상 언어별 음소결합확률모델을 생성하는 훈련과정, 훈련과정에서 생성된 음소결합확률모델로부터 확률값을 계산하여 인식결과를 출력하는 식별과정으로 구성된다. 본 논문에서는 음소결합확률모델로부터 우도를 계산할 때 정보이론(Inforamtion Theory, Shannon and Weaver, 1949)을 이용하여 가중치를 적용하는 방법을 제안하였다. 시스템의 훈련 및 실험에는 OGI 11개국어 전화음성 corpus (OGI- TS)를 사용하였으며, 음소인식기는 HTK를 이용하여 구현하였고 음소인식기 훈련에는 NTIMIT 전화음성 DB를 이용하였다.

실험결과 11개국어를 대상으로 45초 길이의 음성에 대해서 평균 74.1%, 10초 길이의 음성에 대해서는 평균 57.1%의 인식률을 얻을 수 있었다.

1. 서론

음성으로부터 중요한 정보를 얻어내고 처리하는 기술들은 수 십년 전부터 연구되어 오고 있으며 그 중 일부는 일상생활에 적용할 수 있는 수준에 이르러 현재 우리 주변에서도 흔히 접할 수 있고 많은 편리함을 제공하고 있다. 음성언어를 처리하는 분야들에는 음성인식, 화자인식, 음성합성, 음성코딩, 언어식별 등 다양한 분야가 있는데 이 중 언어식별분야는 다른 분야에 비해 국내에서는 상대적으로 소홀히 다루어졌던 분야이다.

언어식별은 입력음성이 어느 나라 언어인지를 식별하는 기술로써 음성인식 기술의 발전과 세계화 추세에 더불어 주목받고 있다. 언어식별시스템은 국제 전화망에 적용되기 위한 목적으로 주로 연구되며, 다국어 자동번역시스템의 전단부, 다국어로 구성된 음성정보검색 등 다양한 분야에서 소요되는 필수기술로 여겨지고 있다. 현재 EU로 통합된 유럽을 비롯하여 미국 등 선진국에서는 활발하게 연구가 진행되고 있다.

언어에 포함된 많은 정보들 중에서 언어식별에 중요한 요소는 크게 네 가지로 나누어 볼 수 있다. 첫 번째는 음운론적 요소로서 각 언어들은 각기 고유하며 서로 다른 음소들과 음소의 빈도 그리고 음소의 배열을 가지고 있다는 것이다. 두 번째는 형태론적 요소로서 각 언어는 단어의 형태나 단어를 구성하는 방법이 다르다는 것이다. 세 번째로는 구문론적 요소로서 각 언어는 문장의 구성 형태가 다르다는 것이다. 즉 어떤 언어들에 똑같은 철자의 단어가 있더라도 그 단어들의 품사가 다르고 따라서 앞 뒤에 올 수 있는 단어들이 다를 수 있다. 마지막으로 운율론적 요소로서 각 언어들은 고유한 특성의 음운의 길이, 피치, 강세 등을 가지고 있다는 것이다[2].

본 논문에서는 이러한 요소들 중에서 현재까지 언어식별에 가장 효율적이고 인식률이 우수한 것으로 알려진 음운론적 요소 즉, 음소배열의 차이를 이용하여 언어식별시스템을 구성하고 실험하였다.

본 논문의 구성은 2장에서 본 논문에서 구현한 언어식별 시스템의 전체적인 구성과 구성요소들을 살펴보고 3장에서는 본 논문에서 제안한 정보이론을 이용한 가중치 적용방법에 대해 설명하고 4장에서는 실험내용과 그 결과를 보인 후, 5장에서 결론을 맺는다.

2. 언어식별시스템 구성

2.1 언어식별시스템 구성도

언어식별 과정은 크게 두 단계로 나누어진다. 첫 번째는 발성음을 음소단위로 인식하여 음소열을 출력해내는 음소인식 단계이고, 두 번째는 생성된 음소열을 이용하여 각 언어의 음소결합확률모델에 대해 확률값을 계산하여 우도를 결정하는 언어식별 단계이다. 시스템 구성도는 아래의 [그림 1]과 같다.

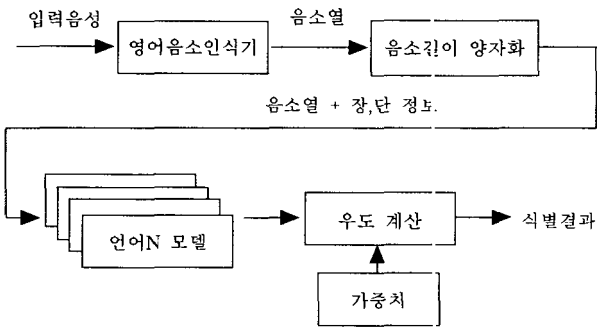

그림 1. 언어식별시스템 구성도

2.2 음소인식기

음소인식부의 구성방법은 여러 가지가 있다. 첫 번째 방법은 식별대상이 되는 모든 언어들에서 대표가 되는 음소들을 뽑아내어 음소집합을 구성하고, 이에 대하여 언어 독립적인 음소인식기를 구현하는 것이다[1]. 하지만 이와 같은 방법은 상당한 언어학적 지식과 기술이 필요하므로 구현하기가 매우 어렵다. 또한 음소인식기를 구현하기 위해서는 각 언어에 대해 음소별로 세그멘트된 훈련자료가 필요하다. 각 언어별로 이러한 음성자료를 구하는 것은 현실적으로 어려우며 새로운 언어의 추가에 대해 유연하지 못한 단점이 된다. 두 번째 방법은 식별대상이 되는 언어들 전부 또는 일부 언어에 대해 음소인식기를 구현한 후 이를 병렬로 인식 처리하는 방법이다[2]. 이와 같은 방법은 인식 및 처리 속도가 느린 단점이 있으며, 첫 번째 방법과 마찬가지로 음소인식 대상 언어별로 음소인식기 훈련에 필요한 음성자료가 준비되어야 하는 단점이 있다.

본 논문에서는 위에 열거한 문제점들을 고려하여 단일 언어 음소인식기를 사용하였다. 훈련에 필요한 음성자료를 쉽게 구하고 접근할 수 있는 영어를 선택하였고, HTK를 이용하여 구현하였다. 인식기 구현에 사용된 특징벡터는 에너지를 포함한 MFCC, Delta, Delta Delta 계수에 대해 각각 13차씩 총 39차를 사용하였다. HMM의 구조는 시작과 끝을 null transition으로 하는 5 state continuous HMM을 사용하였으며 state당 mixture의 개수는 5개로 하였다. 음소인식기 훈련에는 NTIMIT telephone-speech corpus를 사용하였고, shibboleth sentence(사투리)를 제외한 모든 문장을 훈련에 사용하였다. 인식대상이 되는 음소는 NTIMIT에 총 61개로 구분되어 세그멘트되어 있지만, 여기서는 유사음소들을 결합하여 총 48개로 구성하였다[3]. 인식대상이 되는 음소목록은 아래의 [표 1]과 같다.

표 1. 인식대상 음소

종 류	음 소
폐쇄음	b, d, g, p, t, p, k, dx
파찰음	jh, ch
마찰음	s, sh, z, zh, f, th, v, dh
비음	m, n, ng, en
반모음,운음	l, r, w, y, hh, el
모음	iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, er, ax, ix,
기타	sil, epi(삽입 휴지음)

음소인식기의 인식률은 아래의 [표2]와 같다.

표 2. 음소인식률

인식률(%)	N	I	S	D	H
51.01%	51,681	7,146	21,302	4,015	26,364

[표2]에서 N은 총 음소 개수, I는 삽입오류, S는 치환오류, D는 삭제오류, H는 올바르게 인식한 음소의 개수를 의미한다.

2.3 음소길이 양자화

운율론적요소 중의 하나인 음소길이 정보를 이용하는 가장 간단한 방법으로 음소길이 양자화기(Duration Quantizer)를 적용하였다[2]. 음소길이 양자화기는 훈련에 사용된 모든 음성에서 각 음소에 대해 평균길이를 계산하여 테이블로 저장한다. 인식시에는 음소인식기의 출력결과 음소에 대해 이미 계산된 평균값보다 그 길이가 긴지 짧은지를 판단한다. 음소결합확률모델 생성에는 이렇게 음소길이 양자화기를 거쳐 장단여부가 태깅된 음소가 사용된다.

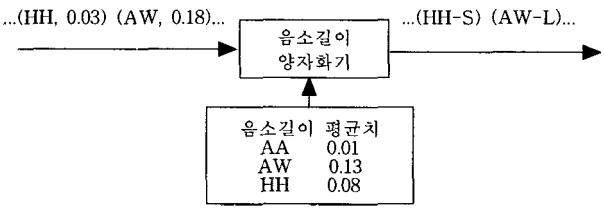

그림 2. 음소길이 양자화

2.4 음소결합확률모델

식별대상이 되는 언어별 음소배열의 특징, 즉 음소결합확률모델은 언어식별에 가장 중요한 요소이다. 훈련용 음성자료에 대해 음소인식기의 인식결과로 출력된 음소열을 이용하여, 각 언어별로 음소와 음소열의 출현빈도를 계산하여 음소결합확률모델로 N-gram 언어모

델을 생성한다.

본 논문에서는 바이그램 이상의 연속된 음소열을 사용한 경우가 바이그램을 사용한 경우와 비교하여 그다지 좋은 성능을 볼 수 없었다는 실험결과[2]를 토대로 연속된 두 개 음소의 출현확률을 계산하는 interpolated bigram[4]을 사용하였다. interpolated bigram을 계산하는 확률식은 아래와 같다.

$$\hat{P}(w_t \mid w_{t-1}) = \alpha_2 P(w_t \mid w_{t-1}) + \alpha_1 P(w_t) + \alpha_0 P_0 \quad (1)$$

$$P(w_t \mid w_{t-1}) = \frac{C(w_{t-1}, w_t)}{C(w_{t-1})} \quad (2)$$

위 식에서 $C(w_{t-1}, w_t)$는 음소 w_{t-1}과 w_t가 연속으로 출현한 횟수이고 $C(w_{t-1})$는 음소 w_{t-1}이 출현한 횟수이며, P_0는 인식대상 음소 개수의 역수이다. 위 식에서 α계수 들의 값은 실험적으로 결정하는데 다른 논문들의 실험 결과를 보면 $0.3 < \alpha_1, \alpha_2 < 0.7$ 사이에서 가장 좋은 결과를 얻을 수 있었다고 한다[2]. α의 값은 훈련용 음성자료의 양에 따라 결정될 수 있는데, 음성자료의 양이 증가할수록 높은 차수의 α값이 커지게 된다. 그리고 대체로 식별대상 음성의 길이가 길수록 높은 차수의 α값이 클 때 인식성능이 좋아지는 것을 볼 수 있다.

인식과정에서는 전단부의 음소인식 결과로 생성된 음소열 W=($w_0, w_1, w_2, \cdots$)을 이용하여 훈련된 각 언어모델 λ_l^{BG}에 대한 로그우도를 아래의 식(3)과 같이 계산한다.

$$L(W \mid \lambda_l^{BG}) = \sum_{t=1}^{T} \log \hat{P}(w_t \mid w_{t-1}, \lambda_l^{BG}) \quad (3)$$

최종 언어식별 과정은 각 언어에 대해 위의 식(3)과 같이 계산된 결과값을 최대로 하는 값을 인식된 결과로 출력하게 된다.

$$\hat{l} = \arg\max_l L(W \mid \lambda_l^{BG}) \quad (4)$$

3. 정보이론을 이용한 가중치 적용

3-1 정보이론(Information Theory)

정보이론(Shannon and Weaver, 1949)은 어떤 선택의 결과에 대해 기대되는 정보의 양을 판단할 수 있는 수학적 방법론이다. 동전던지기의 예를 들어보면 그 의미를 쉽게 알 수 있다. 어떤 동전이 앞면이 나올 확률이 99%라고 하면 우리는 이미 그 결과를 충분히 예측할 수 있으므로 실제 결과가 가지는 정보의 양은 매우 적다고 할 수 있다. 다시 말하면 우리가 적게 알수록 정보의 양은 많아지고 이미 많이 알고 있다면 정보의 양이 적어진다는 것이다. 일반적으로 가능한 답 v_i가 확률 $P(v_i)$를 가진다면 실제응답이 가지는 정보의 양은 아래 식(5)와 같다.

$$I(P(v_1), \ldots, P(v_n)) = \sum_{i=1}^{n} -P(v_i) \log_2 P(v_i) \quad (5)$$

이를 동전던지기 경우에 적용하여 보면

$$I(\frac{1}{2}, \frac{1}{2}) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

즉 확률이 반반일 경우에는 정보의 양이 1이되고 만약 한쪽면이 나올 확률이 100%라면 정보의 양은 0이된다.

3-2 가중치 적용

앞에서 설명한 정보이론을 바탕으로 본 논문에서는 우도계산시에 정보양이 많은 쪽에 가중치를 높게 주는 방법을 제안하였다. 가중치를 계산하는 식은 아래 식(6)과 같다.

$$WT_{t,t-1} = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n} \quad (6)$$

$$WT_{t,t-1} = WT(w_t, w_{t-1} \mid \lambda_l, \lambda_m)$$

$$p = P(w_t \mid w_t, \lambda_l), n = P(w_t \mid w_t, \lambda_m)$$

이 가중치를 우도계산식 (3)에 적용하면 아래의 식(7)과 같다.

$$L(W \mid \lambda_l^{BG}) = \sum_{t=1}^{T} WT_{t,t-1} * \log \hat{P}(w_t \mid w_{t-1}, \lambda_l^{BG}) \quad (7)$$

가중치를 적용하는 식에서 볼 수 있는 것처럼 이 방법은 언어모델 각 두 개씩 짝을 지어 계산하게 되므로 식별과정도 두 개씩 짝을 만들어 토너먼트 방식으로 진행하였다.

4. 실험결과

4.1 OGI 전화음성 DB

본 논문에서 언어식별실험에 사용한 음성 DB는 Oregon Graduate Institute Multi-Language Telephone Speech(OGI-TS) corpus이다[5]. OGI-TS는 총 11개국어로 구성되어 있다. 미국내에 거주하는 각 언어별 원어민을 대상으로 미국내의 전화를 통해 수집된 전화음성 Corpus이며, 미리 작성된 질문에 각 실험대상 화자들이 응답하는 내용을 녹음한 것이다. 질문은 고정된 어휘로 응답을 유도하는 네가지 항목과(예, 일부터 십까지 발음하시오), 자유로운 어휘의 응답을 유도하는 여섯가지 항목으로(예, 당신의 고향의 기후에 대해 말하시오) 구성되어 있다. OGI-TS의 구성은 아래의 [표 3]과 같다.

표 3. OGI-TS Corpus 구성

Language	Initial Training		Development Test		Extended Traing		Final Test	
	M	F	M	F	M	F	M	F
English	33	17	14	6	72	30	16	4
Farsi	39	10	15	4	8	1	18	2
French	40	10	15	5	11	2	12	8
German	25	25	11	9	10	5	15	5
Hindi	47	3	13	4	25	11	14	6
Korean	32	17	18	2	3	2	15	5
Japanese	30	20	15	5	1	0	11	8
Mandarin	34	15	14	6	8	8	10	10
Spanish	34	16	16	4	14	5	11	8
Tamil	43	7	17	3	20	2	19	1
Vietnam	31	19	16	4	11	6	13	7

훈련에 사용한 음성은 훈련 set과 개발실험 set에 포함된 자유로운 어휘의 응답을 유도하는 질문에 대한 다양한 길이(10초~60초)의 응답음성 전체를 사용하였고, 실험에는 최종 실험 set에 포함된 음성 중 자유주제 발화만을 사용하였다. 최종 실험 set에 포함된 45초 길이의 자유주제 발화음성을 각 언어별로 20개씩 사용하였고, 10초 길이의 음성은 45초 음성에서 10초 단위로 3개씩 잘라내어 만든 60개를 사용하였다.

4.2 결과

실험결과는 아래의 [표4]와 같다. 실험은 Baseline system에 대한 것과 본 논문에서 지안한 정보이론(Information Theory)을 이용한 가중치 적용방법을 비교하여 수행하였다. 실험결과에서 볼 수 있듯이 제안한 가중치 적용방법을 이용할 경우 인식률이 향상되는 것을 볼 수 있었다.

표 4. 인식결과

Language	Baseline		가중치 적용	
	45s	10s	45s	10s
English	16	41	14	33
Farsi	16	39	16	33
French	15	42	16	47
German	16	33	17	36
Hindi	12	26	13	30
Japanese	14	30	14	32
Korean	13	32	14	35
Mandarin	15	35	14	29
Spanish	12	26	13	30
Tamil	18	47	18	48
Vietnam	12	20	14	24
합계	159	371	163	377
인식률(%)	72.3	56.2	74.1	57.1

[표5]와 [표6]에는 실험결과 Confusion Matrix를 만들어 보았다. 실험에 사용한 음성의 개수가 좀 더 많다면 이와 같은 결과에서도 언어식별에 유용한 특징을 찾아볼 수 있을 것으로 기대된다.

표 5. 45초 음성 Confusion Matrix

	En	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi
En	14	0	0	2	0	0	0	2	1	0	1
Fa	0	16	1	0	0	0	0	0	0	1	2
Fr	0	0	16	1	0	1	0	0	1	0	1
Ge	1	0	1	17	1	0	0	0	0	0	0
Hi	1	3	2	0	13	0	0	0	0	1	0
Ja	0	1	2	1	0	14	0	0	1	0	1
Ko	1	0	2	1	0	1	14	0	0	0	1
Ma	0	0	1	3	0	1	0	14	0	1	0
Sp	2	0	2	1	0	0	1	0	13	1	0
Ta	0	1	0	0	0	0	0	0	1	18	0
Vi	1	0	0	0	1	1	1	0	1	1	14

표 6. 10초 음성 Confusion Matrix

	En	Fa	Fr	Ge	Hi	Ja	Ko	Ma	Sp	Ta	Vi
En	33	3	0	4	0	1	4	7	1	2	5
Fa	0	33	6	1	4	0	6	0	1	1	8
Fr	1	0	47	2	1	2	0	1	2	1	3
Ge	5	1	5	36	2	4	2	0	4	0	1
Hi	0	6	8	2	30	2	4	1	1	3	3
Ja	0	3	5	2	2	32	0	1	7	2	6
Ko	2	4	5	4	1	4	35	0	3	1	1
Ma	0	3	5	4	0	2	3	29	6	4	4
Sp	4	1	6	3	3	2	3	0	30	7	1
Ta	0	1	0	1	0	2	2	0	2	48	4
Vi	2	4	0	2	2	2	4	5	5	10	24

5. 결론

본 논문에서는 음성인식의 다양한 분야 중에서 상대적으로 국내에서 관심이 적었던 언어식별 분야에 대해 연구하였고, 음소인식기와 음소결합확률모델을 이용하여 언어식별시스템을 구현하고 실험하였다. 간단한 바이그램 언어모델을 이용하였는데도 상당한 인식률을 볼 수 있었고, 정보이론을 이용한 가중치 적용시 인식률이 향상되는 것을 볼 수 있었다. 언어식별에 유용한 특징들을 추출하고 이용하는데 좀더 많은 연구를 추진하고, 실용화 가능한 시스템을 구현할 수 있도록 인식률 향상을 위한 연구를 수행할 예정이다.

언어식별기술은 다국어 음성정보처리에 필수적인 분야이다. 국제화 시대를 논하는 것은 진부한 표현일 것이고 언어식별 시스템의 필요성이 날로 증대할 것임은 자명한 일이다. 앞으로 언어식별 분야에 대한 관심이 증대되기를 기대한다.

6. 참고문헌

[1] T. J. Hazen and V. W. Zue, "Segment-Based Automatic Language Ide ntification", Journal of the Acoustical Society of America, Vol. 101, No. 4, pp. 2323-2331, April 1997.

[2] M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", in IEEE Trans. Speech and Audio Proc., SAP-4(1), January 1996.

[3] Y. S. Yun, "Performance Improvement of CSR Using A Segmental-Feature HMM, , Ph.D thesis, Korea Advanced Institute of Science & Technology, November 2000.

[4] F. Jelinek, "Self-organized language modeling for speech recognition", in Readings in Speech Recognition, A. Waibel and K.-F. Lee, Eds. Palo Alto, CA: Morgan Kaufmann, 1990, pp. 450-506.

[5] Y. K. Muthusamy, R. A. Cole and B. T. Oshika "The OGI multi-language telephone speech corpus", in Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP 92), Alberta, October 1992.