한국음향학회지 제31권 제3호 pp. 170~178 (2012)
The Journal of the Acoustical Society of Korea Vol.31, No.3 (2012)

음소 결정트리의 노드 분할을 위한 임계치 자동 결정 알고리즘

The Automated Threshold Decision Algorithm for Node Split of Phonetic Decision Tree

김범승 · 김순협*

(Beom-Seung Kim and Soon-Hyob Kim*)

코레일 정보기술단, *광운대학교 컴퓨터공학과

(접수일자: 2012년 2월 24일; 수정일자: 2012년 3월 16일; 채택일자: 2012년 4월 1일)

초 록: 본 논문에서는 코레일에서 운영중인 640개 기차역명의 음소기반의 음성인식을 위하여 트라이폰 단위의 음소 결정트리 구축 시 노드 분할 과정에서 사용되는 임계치의 결정에 있어 통계적 기법인 상관관계 분석과 회귀분석을 활용하여 군집화율을 추정하고 이를 이용한 평균 군집화율에 따른 임계치의 값에 의해 자동으로 결정하는 방법을 제안하였다. 제안된 방법의 유효성 검증을 위한 실험에서 기존의 일괄 적용된 Baseline 보다 1.4~2.3 %의 인식률 향상을 보였다.
핵심용어: 기차역명 음성인식, 음소 결정트리, 유사음소단위, 자동음성인식
투고분야: 음성처리 분야 (2.5)

ABSTRACT: In the paper, phonetic decision tree of the triphone unit was built for the phoneme-based speech recognition of 640 stations which run by the Korail. The clustering rate was determined by Pearson and Regression analysis to decide threshold used in node splitting. Using the determined the clustering rate, thresholds are automatically decided by the threshold value according to the average clustering rate. In the recognition experiments for verifying the proposed method, the performance improved 1.4~2.3 % absolutely than that of the baseline system.
Key words: Speech recognition of train station, Phonetic decision tree, PLU, ASR
ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

전화망을 통한 철도예약서비스 (IVR)[1], 자동티켓 발매기 (ATIM)[2], 역안내서비스 (KIOSK) 등 고객접점의 자동화서비스에 음성인식을 적용하기 위하여 가장 먼저 고려해야 할 대상은 「기차역에 대한 역명 인식」을 위한 음성인식 DB의 구축이다[3]. 현재 코레일에서 운영하는 광역지하철역을 포함한 640개의 기차역명은 최소 인식 어휘이다. 음성인식을 위한 모델링의 기본단위로 단어, 음절, 음소, PLU (유사음소 단위: Phoneme-Likely Unit) 등 을 사용할 수 있다[3].

***Corresponding author:** 김범승 (bluedav@korail.com)
100-162 서울시 중구 봉래동 2가 122 철도빌딩 425호 코레일
(전화: 02-3149-3421; 팩스: 02-361-8367)

본 논문에서는 640개의 기차역명의 트라이폰 단위의 음소기반의 음성인식을 위하여 46 PLU (표 1의 State [2 ~ 4]의 각 열, sil 포함)를 사용하였다. 하지만 기차역명의 트라이폰 단위의 음성인식을 위한 음향모델링시 훈련데이터의 부족으로 Unseen Data 에 대한 문제가 발생한다. 이러한 문제를 효율적으로 해결하기 위하여 트라이폰 단위의 음소 결정트리 (Phonetic Decision Tree)를 이용한 상태공유 방법을 사용한다[4,8]. 이 방법은 결정트리의 분류와 예측으로 훈련 데이터에서 나타나지 않은 모델의 합성을 가능하게 하고 결정트리 기반의 상태공유를 위한 노드 분할 과정과 모델 선택 과정을 통해 모델의 복잡성을 완화시키고 한정된 훈련 데이터로부터 강건한 모델 파라

미터 추정을 가능하게 하여 필요한 파라미터 양과의 균형을 유지할 수 있는 장점을 가지고 있다[4-7]. 음소 결정트리에서는 중심음소를 기준으로 음성학적 질의에 의해 새롭게 생성된 음향모델은 군집화된 어느 하나의 덩어리에 포함되어 상태를 공유하며 미지의 음소에 대하여 군집화된 대표 상태를 공유하게 됨으로써 인식률의 향상을 가져올 수 있다. 이러한 군집화의 정도의 결정에 영향을 줄 수 있는 임계치의 설정은 대부분 실험치에 의하여 일괄 적용하거나 음소 단위로 가변적으로 적용한다. 하지만 실험치에 의해 이러한 부분을 정확히 가늠하기는 쉽지가 않다[4-7]. 음소의 개수에 대한 상태별 군집화 정도를 판단해야 하기 때문이다. 또한 이것이 반드시 인식률 향상에 가져온다고 볼 수 없다. 군집화의 정도가 너무 크거나 너무 작으면 변별력이 떨어져 인식률이 오히려 감소할 수 있다. 따라서 본 논문에서는 이러한 임계치의 자동결정을 위하여 앞서 언급한 음소별 상태별 군집화율에 대하여 PLU 빈도수 와 군집화율의 상관관계를 분석하고 이를 바탕으로 회귀분석을 통하여 도출된 회귀식에 의해 기차역명에서의 음소의 빈도수에 대한 자료를 바탕으로 군집화율을 추정하고 추정된 군집화율에 따라 임계치를 자동으로 결정하도록 한다. 이러한 방법이 기존 수작업 임계치 결정방법보다 제안된 임계치 자동결정알고리즘 에서의 인식률이 크지는 않지만 1.4~2.3 %의 인식률의 향상을 보임을 알 수 있다. 이는 적어도 제안된 임계치 자동결정알고리즘이 실험에 의하여 수작업으로 임계치를 결정하는 방식 보다는 효율적이며 유효성이 확인할 수 있다. 본 논문의 구성은 Ⅰ장 서론에 이어, Ⅱ장에서는 이론적 배경에 대하여 살펴보고, Ⅲ장에서는 제안하는 방식에 대하여 알아보고, Ⅳ 실험을 통하여 제안하는 방식의 알고리즘의 유효성을 확인하고, Ⅴ장에서는 결론 및 향후 계획에 대하여 이야기한다.

Ⅱ. 이론적 배경

한 음소에 대한 결정트리는 어느 모델들이 특정 문맥내에서 사용될지를 결정해줌으로써 하나의 트라이폰이 주어지면 루트노드에서부터 출발하여 현재의 문맥에 관한 음성학적 질의(「왼쪽 또는 오른쪽 음소가 집합 X의 원소인가?」의 형태)의 결과에 따라 다음노드를 선택하는 방식의 트리순회를 통하여 하나의 모델을 선택한다. 이런 트리는 훈련 데이터에 없던 트라이폰 모델을 합성해 낼 수 있고, 해당 트라이폰 모델의 문맥과 가장 비슷한 단말 트리 노드를 찾아 이와 연관된 공유상태들로 구성 할 수 있다[8-10]. 임의의 S를 HMM 상태들의 집합이라고 하고 $L(S)$를 S의 Log Likelihood라 할 때 집합 S에 있는 모든 상태들이 묶였다는 가정하에 훈련 데이터의 프레임들의 집합인 F에서 생성된다고 하면, 공통 평균 $\mu(S)$와 분산 $\Sigma(S)$를 공유한다[10]. (단, 전이 확률은 무시된다.) 또한 묶인 상태들의 상태별 프레임의 정렬이 바뀌지 않는다고 가정하면 $L(S)$에 대해 수식 (1)과 같이 쓸 수 있다.

$$L(S)=\sum_{f\in Fs\in S}\sum\log(P(O_t;\mu(S),\Sigma(S)))r_s(O_f) \quad (1)$$

$r_s(O_f)$는 상태 r_s에 의하여 생성되는 관측 프레임 O_f의 사후확률이다. 만일 출력확률밀도함수가 가우시안식이면 수식 (2)와 같이 쓸 수 있다.

$$L(S)=-\frac{1}{2}(\log[(2\pi)^n|\Sigma(S)|]+n)\sum_{f\in Fs\in S}\sum r_s(O_f) \quad (2)$$

n은 데이터의 차수를 나타낸다. 그러므로 전체 데이터 집합의 로그 확률은 상태들의 분산 $\Sigma(S)$와 전체 상태 점유인 $\sum_{f\in Fs\in S}\sum r_s(O_f)$만으로 계산할 수 있게 된다. 상태들의 평균과 분산으로부터 앞부분을 계산할 수 있으며, 상태 점유수는 이전 단계의 Baum-Welch 재추정하는 동안 저장될 수 있다. 질의 q에 의해 두 하위 집합 $S_y(q)$와 $S_n(q)$를 나누는 상태들 S와 함께 주어진 노드일 경우, 그 노드는 다음 수식 (3)을 최대화하는 질의 q^*를 사용하여 분할한다.

$$\triangle L_q=L(S_y(q))+L(S_n(q))-L(S) \quad (3)$$

수식 (3)의 $L(S_y(q^*))$는 음성학적 질문 q에 대한 yes로 응답한 $node$의 로그유사도 (Log Likelihood)이고, $L(S)$는 분할하기전의 파티션 전체의 로그유사

도이다. $\triangle L_q$가 최대가 되는 지점에서 질문과 분할 노드를 선택하며 여기에서 가장 적합한 질의를 사용했을 때 로그 유사도의 증가량이 기준 문턱치 보다 높은 경우에만 분할을 하며 그 이득이 특정 임계치가 될 때 까지 이러한 과정을 반복한다. 본 논문에서는 이러한 임계치의 결정을 위하여 통계적 기법인 상관관계 분석과 회귀분석을 활용하고자 하였다[11].

III. 제안하는 방법

3.1. 음소의 빈도수에 따른 군집화율

본 논문에서는 표 1과 같이 기차역명에서의 음소의 빈도수 (P_Count), 훈련 데이터의 음소의 상태별 점유 빈도수 (S_Count), 음소 결정트리에서의 임계치 증가에 따른 음소별 군집화율의 평균치 (C_Rate)를 가지고 그 상관관계를 분석하고 각 빈도수에 따른 군집화율을 추정하고자 하였다. 그림 1에서 알 수 있듯이 640개의 기차역명에서 음소의 빈도수가 높을수록 군집화율이 감소하고 있는 것을 알 수 있다. X축의 음소별 빈도수 대비 Y축의 군집화율의 상관관계를 분석해보면 State [2]~State [4]의 총 빈도수 대비 군집화율에서 Pearson 상관계수는 -0.501의 부의 상관관계를 가진다. 이는 State Tying 시 음소의 빈도수가 적은 경우에 Unseen Data에 대하여 질의에 의해 더 많은 상태공유를 가진다는 것으로 이해할 수 있다. 그리고 훈련 데이터의 음소의 상태별 점유 빈도수와 앞서 분석한 음소의 빈도수와의 상관관계를 보면

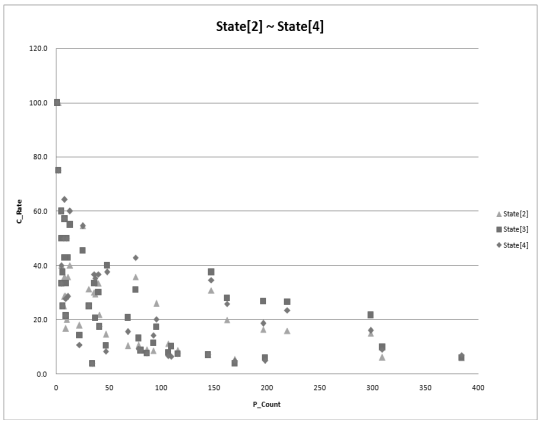

그림 1. 음소의 출현 빈도수에 따른 군집화율
Fig. 1. Phoneme Frequency and Clustering Rate.

Pearson 상관계수 0.949의 정의 강한 상관관계를 가지고 있음을 알 수 있다. 이는 기차역명의 음소의 출현 빈도수가 클수록 훈련데이터의 음소의 상태별 점유 빈도수가 커지고 있는 것으로 이해할 수 있다. 또한 음소의 상태별 점유 빈도수와 군집화율의 상관관계를 보면 Pearson 상관계수 -0.550으로 음소별 상관관계와 같이 부의 상관관계를 가지고 있음을 알 수 있다. 이 또한 훈련 데이터의 음소의 상태별 점유 빈도수가 커질수록 군집화율이 감소하는 것으로 이해할 수 있다. 위에서 살펴본 바와 같이 기차역명의 음소의 빈도수와 훈련 데이터의 음소의 상태별 점유 빈도수는 군집화율에 영향을 주고 있음을 알 수 있으며 상관계수는 모두 0.01 수준에서 양쪽에서 유의하였다. 따라서 음소의 빈도수, 훈련 데이터의 음소의 상태별 점유 빈도수에 따른 군집화율에 대하여 회귀분석하게 되면 다음과 같은 회귀모형을 고려할 수 있다. Model1의 경우는 음소의 빈도수만을 가진 단순 회귀모형, Model2의 경우에는 훈련 데이터의 음소의 상태별 점유 빈도수를 가진 단순 회귀모형, Model3의 경우에는 Model1과 Model2의 변수를 모두 포함한 다중 회귀모형이다. 표 2를 살펴보면 분산분석 (ANOVA) 결과 Model1과 Model2의 경우 P값이 0.000으로 5 % 신뢰수준일 때의 P값의 임계치 0.05보다 작기 때문에 귀무가설을 기각하여 편회귀계수는 0이 아니라고 할 수 있다. 하지만 Model3의 경우 (t 검정에서) 훈련 데이터의 음소의 상태별 점유 빈도수는 P값이 0.001로 5 % 신뢰수준일 때의 P값의 임계치 0.05보다 작기 때문에 유의하였으나 음소의 빈도수는 P값이 0.366으로 5 % 신뢰수준일 때의 P값의 임계치 0.05보다 크기 때문에 유의하지 않았다. 또한 다중 공선성을 진단할 수 있는 통계량인 Tolerance (공차한계)의 값이 0.100, VIF (Variance Inflation Factor: 분산팽창계수)의 값이 10.048로 독립변수간의 약간의 다중 공선성의 문제가 있음을 알 수 있다. 따라서 Model3은 적합하지 않다. 본 논문에서는 Model1과 Model2의 회귀모형을 고려하였다. 회귀분석에 의해 결정된 Model1과 Model2의 수식은 (4), (5)와 같으며 각각 음소의 빈도수 (P_Count), 상태별 점유빈도수 (S_Count)가 1단위 증가할 때 군집화율 C (C_Rate)가 0.116, 0.038 감소함을 의미하며 상수 36.7, 40.2는 절

표 1. 상태 점유빈도수에 따른 군집화율
TABLE 1. Frequency of State Occupancy and Clustering Rate.

Phone	State [2]	P_ Count	S_ Count	C_ Rate	State [3]	P_ Count	S_ Count	C_ Rate	State [4]	P_ Count	S_ Count	C_ Rate
ㄱ	g [2]	169	780	5.4	g [3]	169	780	3.8	g [4]	169	780	4.6
ㄱf	gq [2]	92	420	8.6	gq [3]	92	420	11.4	gq [4]	92	420	14.3
ㄲ	gg [2]	9	84	28.6	gg [3]	9	84	21.4	gg [4]	9	84	21.4
ㄴ	n [2]	106	540	11.1	n [3]	106	540	7.8	n [4]	106	540	6.7
ㄴf	nq [2]	309	792	6.1	nq [3]	309	792	9.8	nq [4]	309	792	9.1
ㄷ	d [2]	115	408	8.8	d [3]	115	408	7.4	d [4]	115	408	7.4
ㄸ	dd [2]	6	48	37.5	dd [3]	6	48	37.5	dd [4]	6	48	25
ㄹ	r [2]	34	312	3.8	r [3]	34	312	3.8	r [4]	34	312	3.8
ㄹf	l [2]	41	276	21.7	l [3]	41	276	17.4	l [4]	41	276	17.4
ㅁ	m [2]	78	456	10.5	m [3]	78	456	13.2	m [4]	78	456	9.2
ㅁf	mq [2]	68	348	10.3	mq [3]	68	348	20.7	mq [4]	68	348	15.5
ㅂ	b [2]	80	348	8.6	b [3]	80	348	8.6	b [4]	80	348	8.6
ㅂf	bq [2]	11	84	35.7	bq [3]	11	84	42.9	bq [4]	11	84	28.6
ㅃ	bb [2]	1	12	100	bb [3]	1	12	100	bb [4]	1	12	100
ㅅ	s [2]	198	612	5.9	s [3]	198	612	5.9	s [4]	198	612	4.9
ㅆ	ss [2]	31	96	31.3	ss [3]	31	96	25	ss [4]	31	96	25
ㅇ	nx [2]	384	1212	5.9	nx [3]	384	1212	5.9	nx [4]	384	1212	6.9
ㅈ	z [2]	144	516	7	z [3]	144	516	7	z [4]	144	516	7
ㅉ	zz [2]	6	48	25	zz [3]	6	48	25	zz [4]	6	48	25
ㅊ	c [2]	109	468	10.3	c [3]	109	468	10.3	c [4]	109	468	6.4
ㅋ	k [2]	5	36	33.3	k [3]	5	36	33.3	k [4]	5	36	33.3
ㅌ	t [2]	22	168	17.9	t [3]	22	168	14.3	t [4]	22	168	10.7
ㅍ	p [2]	47	288	14.6	p [3]	47	288	10.4	p [4]	47	288	8.3
ㅎ	h [2]	86	396	9.1	h [3]	86	396	7.6	h [4]	86	396	7.6
ㅏ	a [2]	298	1164	14.9	a [3]	298	1164	21.6	a [4]	298	1164	16
ㅏ:	aa [2]	7	72	25	aa [3]	7	72	50	aa [4]	7	72	50
ㅑ	ja [2]	37	204	29.4	ja [3]	37	204	20.6	ja [4]	37	204	35.3
ㅓ	v [2]	196	516	16.3	v [3]	196	516	26.7	v [4]	196	516	18.6
ㅔ	e [2]	95	624	26	e [3]	95	624	17.3	e [4]	95	624	20.2
ㅕ	jv [2]	75	252	35.7	jv [3]	75	252	31	jv [4]	75	252	42.9
ㅖ	je [2]	13	120	40	je [3]	13	120	55	je [4]	13	120	60
ㅗ	o [2]	219	948	15.8	o [3]	219	948	26.6	o [4]	219	948	23.4
ㅗ:	oo [2]	10	60	20	oo [3]	10	60	50	oo [4]	10	60	50
ㅘ	wa [2]	48	240	40	wa [3]	48	240	40	wa [4]	48	240	37.5
ㅚ	we [2]	9	108	16.7	we [3]	9	108	33.3	we [4]	9	108	27.8
ㅛ	jo [2]	25	132	54.5	jo [3]	25	132	45.5	jo [4]	25	132	54.5
ㅜ	u [2]	162	816	19.9	u [3]	162	816	27.9	u [4]	162	816	25.7
ㅜ:	uu [2]	5	60	40	uu [3]	5	60	60	uu [4]	5	60	40
ㅝ	wv [2]	36	180	30	wv [3]	36	180	33.3	wv [4]	36	180	36.7
ㅟ	wi [2]	2	24	100	wi [3]	2	24	75	wi [4]	2	24	75
ㅠ	ju [2]	8	84	28.6	ju [3]	8	84	42.9	ju [4]	8	84	64.3
ㅡ	y [2]	40	180	33.3	y [3]	40	180	30	y [4]	40	180	36.7
ㅢ	yi [2]	5	60	40	yi [3]	5	60	50	yi [4]	5	60	60
ㅣ	i [2]	147	624	30.8	i [3]	147	624	37.5	i [4]	147	624	34.6
ㅣ:	ii [2]	8	84	35.7	ii [3]	8	84	57.1	ii [4]	8	84	57.1

표 2. 회귀분석 표
TABLE 2. Table of Regression.

Model1		R	R Square[1)]	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson[2)]
		0.501	0.251	0.245	18.16081	1.807
		Sum of Squares	df	Mean Square	F	P (=Sig.)
	Regression	14688.790	1	14688.790	44.536	0.000
	Residual	43865.412	133	329.815	-	-
	Total	58554.202	134	-	-	-
		Unstandized Coefficients		Standized Coefficients	t	P (=Sig.)
		B	Std. Error	Beta		
	(Constant)	36.687	2.089	-	17.563	0.000
	P_Count	-0.116	0.017	-0.501	-6.674	0.000
Model2		R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
		0.550	0.302	0.297	17.52813	1.883
		Sum of Squares	df	Mean Square	F	P (=Sig.)
	Regression	17691.923	1	17691.923	57.584	0.000
	Residual	40862.279	133	307.235	-	-
	Total	58554.202	134	-	-	-
		Unstandized Coefficients		Standized Coefficients	t	P (=Sig.)
		B	Std. Error	Beta		
	(Constant)	40.204	2.259	-	17.793	0.000
	S_Count	-0.038	0.005	-0.550	-7.588	0.000
Model3		R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
		0.554	0.306	0.296	17.53976	1.908
		Sum of Squares	df	Mean Square	F	P (=Sig.)
	Regression	17945.303	2	8972.651	29.166	0.000
	Residual	40608.899	132	307.643	Tolerance	VIF
	Total	58554.202	134	-	0.100	10.048
		Unstandized Coefficients		Standized Coefficients	t	P (=Sig.)
		B	Std. Error	Beta		
	(Constant)	40.949	2.405	-	17.023	0.000
	P_Count	0.048	0.053	0.209	0.908	0.366
	S_Count	-0.051	0.016	-0.748	-3.254	0.001
Model4		R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
		0.582	0.339	0.334	0.28741	1.568
		Sum of Squares	df	Mean Square	F	P (=Sig.)
	Regression	5.627	1	68.116	68.116	0.000
	Residual	10.986	133	0.083	-	-
	Total	16.613	134	-	-	-
		Unstandized Coefficients		Standized Coefficients	t	P (=Sig.)
		B	Std. Error	Beta		
	(Constant)	1.538	0.037	-	41.500	0.000
	S_Count	-0.001	0.000	-0.582	-8.253	0.000

1) R Square : 종속변수의 총변동 중 독립변수에 의해 설명된 비율로 $0 \le R^2 \le 1$ 범위를 가짐

2) Durbin-Watson : $DW = \sum_{i=2}^{n}(e_i - e_{i-1}) / \sum_{i=1}^{n} e_i^2$ 로 일반적으로 그 값이 2 초과시 자기상관이 존재함

편을 나타낸다. 또한 수식 (6)의 경우에도 같은 의미를 가진다.

$$Model1:\ C = 36.7 - (0.116 \times P\_Count) \quad (4)$$

$$Model2:\ C = 40.2 - (0.038 \times S\_Count) \quad (5)$$

표 2에서와 같이 Model2의 R Square는 0.302로 Model1의 R Square의 0.251 보다 0.051 더 높으므로 Model2의 독립변수가 종속변수에 대하여 설명력이 더 큼을 알 수 있다. 본 논문에서는 Model2의 경우를 선택하기로 한다. Model2의 종속변수인 C_Rate에 대하여 Log10 함수로 변환하여 회귀분석을 하게 되면 변환된 모형인 Model4의 R Square는 0.339 이다. 이는 변환하지 않은 모형인 Model2의 0.302에 비해 큰 값이며, 분산분석 결과인 F 값 또한 68.116으로 Model2의 57.584 보다 큰 값이다. 따라서 변환한 모형이 변환하지 않은 모형보다 종속변수에 대한 설명력이 크다고 할 수 있다. 또한 그림 2에서 알 수 있듯이 Model4의 히스토그램은 Model2의 잔차의 도표에 비해 훨씬 근접해 있으며 정규확률산점도의 경우에도 Model2의 산점도에 비해 45도선에 거의 근접해 있음을 보여주고 있다. 그리고 종속변수는 Model2의 산점도에 비해 예측값의 크기에 관계없이 잔차의 대부분이 일정한 범위 내에서 균등하게 분포되어 있다. 따라서 종속변수인 C_Rate에 Log 값을 취하여 변환한 모형이 변환하지 않은 모형보다 더 좋은 결과를 나타내고 있음을 알 수 있다. Model4의 수식은 (6)과 같다.

$$Model4:\ C = 1.538 - (0.001 \times S\_Count) \quad (6)$$

3.2. 군집화율에 따른 임계치 결정

일괄 적용된 임계치 증가에 따른 인식률을 보면 Th (Threshold) 50.0에서 95.2 %, Th 100.0에서 95.3 %, Th 200.0에서 95.8 %, Th 300.0에서 94.9 %, Th 400.0에서 95.0 %, Th 500.0에서 94.4 %로 Th 200.0에서 95.8 % 최고치를 가지며 계속 증가 할수록 감소 또는 소폭 반등하고 있음을 알 수 있다. 이는 반드시 임계치의 값이 높다고 인식률이 높은 것은 아님을 알 수 있으며 적정한 수준 (실험치)에서 임계치를 결정해야함을 알 수 있다. 하지만, 일괄 적용된 임계치 보다는 음소의 상태별 임계치를 적용하는 것이 최적의 선택이나 이 또한 많은 실험과 시행착오를 통해서만 결정이 가능하며 많은 시간이 필요이다. 따라서 적정수

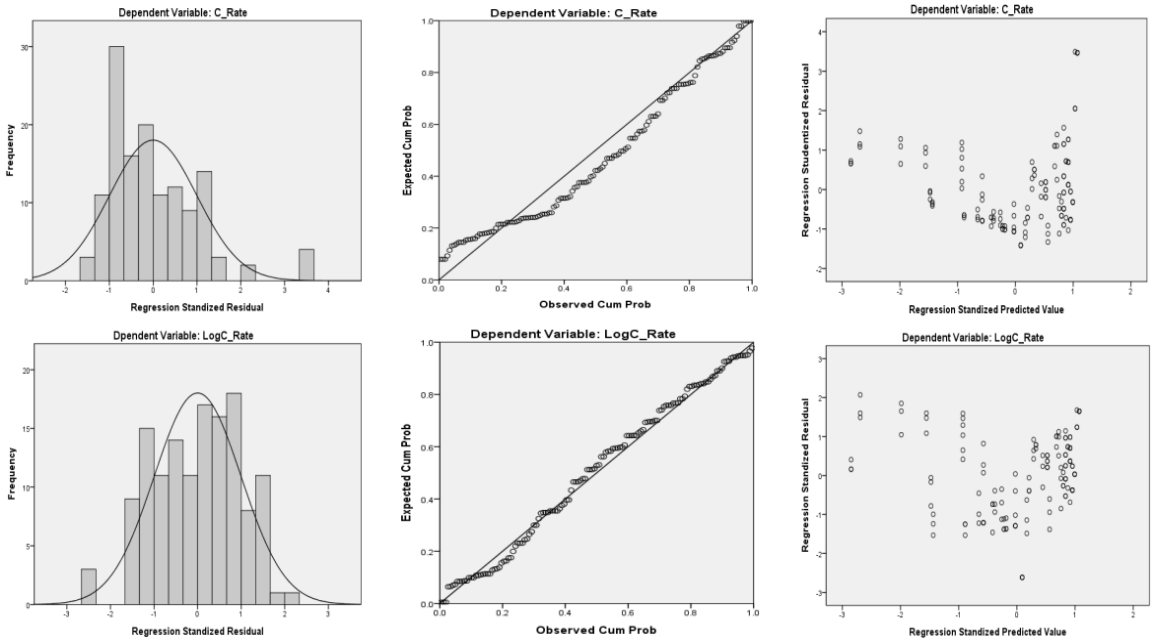

그림 2. 잔차 그래프
Fig. 2. Graph of Residual.

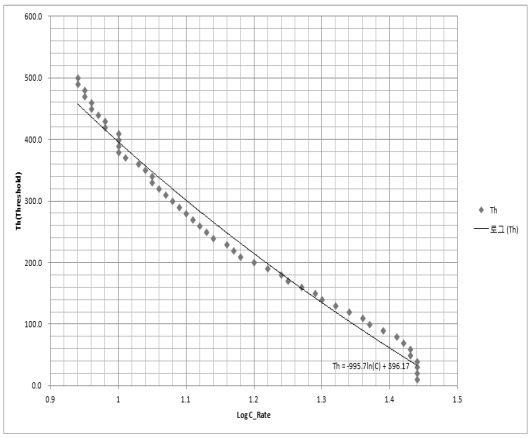

그림 3. 군집화율에 따른 임계치
Fig. 3. Clustering Rate and Threshold.

준의 군집화율에서 임계치를 결정하는 것이 중요하다. 그림 3에서 그래프를 살펴보면 X축의 Log를 취한 평균 군집화율 LogC_Rate와 Y축의 임계치 Th의 상관관계를 분석해보면 Pearson 상관계수 -0.957의 부의 강한 상관관계를 가진다. 이는 LogC_Rate가 증가할수록 거의 선형적으로 Th가 감소하고 있음을 의미한다. 본 논문에서는 임계치별 평균 군집화율에 따른 Th 결정을 최적이라고 가정하고 이를 음소의 상태별 임계치로 사용하고자 하였다. $Th = -995.7 \times \ln(C) + 396.17$은 로그를 적용한 추세선의 수식이다. 여기서 C (X축의 값)는 로그를 취한 군집화율을 의미하며 Th (Y축의 값)는 결정되는 임계치의 값을 의미한다. 따라서 앞서 설명한 Model4의 수식 (6)에서 상태별 빈도수에 따라 결정된 군집화율인 C가 입력값이 되며 이에 따라 Th값 (임계치)이 결정된다. 또한 실험을 통하여 기존 일괄 임계치 적용방법 보다 유효성이 있음을 확인하였다.

그림 4는 $State[2] \sim State[4]$ 까지의 45개의 PLU (sil은 제외)에 대한 Th를 결정하기 위한 알고리즘의 *PSEUDO CODE* 이다. $State[2]$에서 $j = 1$일 때 군집화율 C는 절편 1.538에서 기울기 0.001과 $State[2]$에서의 상태별 점유 빈도수 S_1 (첫번째 PLU에 대한 상태별 점유 빈도수)을 곱한 값을 빼서 결정된다. 이렇게 계산된 군집화율 C가 0.94 이하일 경우에는 Th는 500.0으로 고정하고 1.44 이상일 경우에는 10.0으로 고정한다. 그렇지 않을 경우에는 그림 3의 Th 결정수식인 추세선에 해당하는 로그 수식을 이용하여 앞서

```
FUNCTION(State[i])
FOR  j = 1  TO  45
    C = 1.538 − (0.001 × S_j);
    IF  C ≤ 0.94  THEN
        Th = 500.0;
    ELSE IF  C ≥ 1.44  THEN
    Th = 10.0;
    ELSE
        Th = −995.7 × ln(C) + 396.17;
        /*추세선 위의 값*/
    PRINT(State[i], P[j], Th);
    j = j + 1;
END
```

그림 4. 알고리즘 의사코드
Fig. 4. Pseudo Code for Algorithm.

계산된 군집화율 C를 입력값으로 하여 계산되는 값을 임계치로 결정한다. 결과는 $State[2]$, $P[1]$: 첫 번째 PLU, Th:임계치 순으로 출력한다. $State[2]$에 대하여 $P[45]$ (45번째 PLU) 까지 반복수행하며 State [3], State [4]에 대해서도 같은 과정을 수행한다.

IV. 실험 및 고찰

음소 결정트리에서의 효율적인 임계치 결정을 위해 제안한 임계치 자동 결정알고리즘의 유효성을 확인하기 위하여 각 Model 별 구성을 표 3과 같이 4가지의 경우로 구성하여 음향모델을 작성하였고 트라이폰 단위의 학습용 발음열과 발음사전을 구성하여 인식률 실험을 진행하였다[12,13]. 객관적인 실험을 위하여 표준 음성인식기인 HTK (39 MFCC, CHMM, 3 State, 8 Mixture) [14]를 이용하였고 640개의 기차역명의 인식률 평가를 위해 20~30대 남자 40명 과 여자 40명이 조용한 사무실 환경에서 녹음한 총 80명의 화자 중 학습에 참여한 화자 60명 (남: 30, 여: 30)을 제외한 학습에 참여하지 않은 화자 20명 (남: 10, 여: 10)이 녹음한 음성파일 (샘플링 8 kHz, 양자화 16 bit)을 사용하였다. 그리고 회귀모형에 따라서 결정된 음소의 상태별 임계치에 대하여 인식률을 보면 기존 일괄적용을 Baseline (기준선)으로 볼 때 Model1, Model2,

표 3. 인식률
TABLE 3. Recognition Rate (%).

Model	Baseline	Model1	Model2	Model4
Recog.Rate	95.8	97.2	97.8	98.1

Model4에서의 결정된 Th를 사용하는 것이 Baseline 보다 1.4~2.3 %의 수준에서 인식률이 향상되고 있음을 알 수 있다. 이는 추정된 군집화율에 따라 임계치를 결정하는 것이 기존의 방법보다 아주 큰 폭의 차이는 없지만 좀 더 세밀한 임계치를 줄 수 있고 음성인식에 있어서 유효성이 있음을 알 수 있다. 모델에 따른 인식률은 그림 표 3과 같고 Model4에서 98.1 %의 인식률을 보였다.

V. 결론 및 향후계획

본 논문에서는 코레일에서 운영중인 640개 기차역명의 음소기반의 음성인식을 위한 트라이폰 단위의 음향 모델링 시 발생할 수 있는 훈련 데이터의 부족 문제를 해결하기 위하여 음소 결정트리를 이용한 상태공유 방법을 사용하였다. 이를 위한 음소 결정트리 구축시 노드 분할을 위하여 사용되는 임계치의 결정에 있어 통계적 기법인 상관관계 분석과 회귀분석을 활용하여 군집화율을 추정하고 이를 이용하여 임계치를 자동으로 결정하는 방법을 제안하였다. 제안된 방법의 유효성 검증을 위한 인식률 실험에서 기존의 일괄 적용된 Baseline (95.8 %) 보다 1.4~2.3 % (97.2~98.1 %) 향상된 인식률을 보였다. 이러한 통계적 기법을 활용한 방식은 한국의 기차역명의 음성인식을 위한 좋은 토대가 될 것으로 기대된다. 향후 좀 더 확장된 기차역명 (각 지역의 지하철역명이 포함된) 및 지명에 대하여도 이러한 분석을 통하여 좀 더 효율적인 인식방법에 대해 연구할 계획이다.

감사의 글

본 연구는 광운대학교 2012학년도 교내 학술 연구비 지원으로 이루어졌습니다.

참고문헌

1. B. S. Kim, S. H. Kim, “A Study on Realization of Speech Recognition System based on VoiceXML for Railroad Reservation Service,” *Jounal of the Korea Society for Railway*, vol. 14, no. 2, pp. 130-136, 2011.
2. B. S. Kim, S. H. Kim, “A Study on the Speech Recognition for Commands of Ticketing Machine using CHMM,” *Jounal of the Korea Society for Railway*, vol. 12, no. 2, pp. 285-290, 2009.
3. B. S. Kim, S. H. Kim, “A Study on Speech Recognition based on Phoneme for Korean Subway Station Names,” *Jounal of the Korea Society for Railway*, vol. 14, no. 3, pp. 285-290, 2011.
4. A. Lazarides, Y. Normandin, and R. Kuhn, “Improving decision trees for acoustic modeling,” *in Proc. ICSLP*, Philadelphia, October. 1996.
5. D. B. Paul, “Extensions to phone-state decision-tree clustering: single tree and tagged clustering,” *in Proc. ICASSP*, vol. 2, pp. 1487-490, 1997.
6. L. Gu, K. Rose, “Sub-state tying in tied mixture hidden Markov models,” *Proc. IEEE, Acoustics, Speech, and Signal Processing*, pp. 1062-1065, 2000.
7. R. D. R. Fagundes, J. S. Correa, P. Dumouchel, “A New Phonetic model for continuous speech recognition systems,” *Proc. ICSP*, pp. 572-575, 2002.
8. S. J. Young, J. J. Odell, and P. C. Woodland, “Tree-Based State Tying for Hight Accuracy Accoustic Modelling,” *in Proceedings of the Workshop on Human Language Technology*, Plainsboro, NJ, Mar. 1994.
9. J. J. Odell, “The Use of Context in Large Vocabulary Speech Recognition,” PhD's Dissertation, *University of Cambridge*, 1995.
10. T. O. Ann, “A Study on the Optimization of State Tying Acoustic Models using Mixture Gaussian Clustering”, *Jounal of Electronics Engineers of Korea*, vol. 42, no. 6, pp. 167-176, Nov. 2005.
11. 김성호, 최태성, 사회과학을 위한 통계자료분석 (SPSS 11.0활용), *다산출판사*, 2004.
12. L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, 1989.
13. D. Jurafsky and J. H. Martin, Speech and Language Processing, *PrenticeHall (2nd)*, 2008.
14. S. Young, G. Evermmana, M. Gales, T. Hain, et al, “The HTK Book for HTK Version 3.4,” 2006.

저자 약력

▸ 김 범 승 (Beom-Seung Kim)

1999년: 광운대학교 정보통신대학원 컴퓨터공학과 (석사)
2009년: 광운대학교 대학원 컴퓨터공학과 (박사수료)
1995년 ~ 현재: 코레일 정보기술단 영업정보처 차장
〈관심 분야〉 음성인식

▸ 김 순 협 (Soon-Hyob Kim)

1974년: 울산대학교 전자공학과 학사
1976년: 연세대학교 대학원 전자공학과 석사
1983년: 연세대학교 대학원 전자공학과 박사
1979년 ~ 현재: 광운대학교 컴퓨터공학과 교수
〈관심 분야〉 음성인식