한국음향학회지 제22권 제4호 pp. 323~328 (2003)

벡터 회귀 트리를 이용한 한국어 에너지 궤적 생성

Generating Korean Energy Contours Using Vector-regression Tree

이 상 호*, 오 영 환**
(Sangho Lee*, Yung-Hwan Oh**)

*LG전자기술원 모바일 멀티미디어 연구소, **한국과학기술원 전자전산학과 전산학전공
(접수일자: 2003년 4월 7일; 채택일자: 2003년 5월 6일)

본 논문에서는 한국어 TTS 시스템을 위한 에너지 궤적 생성 방법에 대해 설명한다. 에너지 궤적 생성을 위해 스칼라 회귀 트리를 확장한 벡터 회귀 트리를 제안하고 구현하였다. 벡터 회귀 트리는 특징 벡터로부터 목적 벡터를 예측할 수 있으며, 본 연구에서는 각 음소당 10개의 에너지 값을 예측한다. 실험을 위해 500 문장의 문장 코퍼스와 그 문장들을 발성한 음성 코퍼스를 수집하였고, 이중 300 문장을 이용하여 트리들을 학습하고 200 문장에 대해 실험하였다. 에너지 궤적의 예측 정확률을 높이기 위해 배깅 트리 (bagged tree)와 재구축 트리 (born again tree)도 함께 구현한 결과, 원음의 에너지 궤적과 예측된 에너지 궤적간의 상관계수가 0.803으로 기존의 방법보다 더 좋은 결과를 얻을 수 있었다.
핵심용어: 에너지 궤적, 문서 음성 변환 시스템, 벡터 회귀 트리, 운율, 음성 합성
투고분야: 음성처리 분야 (2.4)

This study describes an energy contour generation method for Korean TTS systems. We propose a vector-regression tree, which is a vector version of a scalar regression tree. A vector-regression tree predicts a response vector for an unknown feature vector. In our study, the tree yields a vector containing ten sampled energy values for each phone. After collecting 500 sentences and its corresponding speech corpus, we trained trees on 300 sentences and tested them on 200 sentences. We construct a bagged tree and a born again one to improve the performance of contour prediction. In the experiment, we got a 0.803 correlation coefficient for the observed and predicted energy values.
Keywords: *Energy contour, TTS system, Vector-regression tree, Prosody, Speech synthesis*
ASK subject classification: *Speech signal processing (2.4)*

I. 서 론

운율은 청자로 하여금 발화자의 음성으로부터 특정 단어에 집중하게 한다거나 혹은 단어들간의 군집화를 하게 하여 발화 의미를 해석하는데 도움을 준다. 운율의 구성 요소에는 음의 경계, 길이, 높낮이, 강세가 있으며 문서 음성 변환 시스템은 운율 생성부 내에 에너지 궤적 생성 모듈을 포함하게 된다.

에너지 궤적 생성 방법은 코퍼스에 기반하는 통계 모델이 주류를 이루고 있다. 예를 들어 신경회로망을 이용하는 방법[1], Bagshaw의 방법[2], 다이나미컬 (dynamical) 시스템[3] 등이 있다. 예측 값들은 각각의 방법에 따라 크게 음절 수준[1], 음소 수준[2], 프레임 수준[3]에서 예측된다.

본 논문에서는 표준 스칼라 회귀 트리를 확장한 벡터 회귀 트리를 제안하고 이 트리를 이용하여 한국어 에너지 궤적을 예측하고자 한다. 트리 생성 방법은 트리 확장 단계와 트리 제거 단계로 이루어져 있고, 각각 Chou의 방법과[4] 비용-복합도 제거 (cost-complexity pruning) 방법을[5] 이용하여 구현하였다. 트리 기반 예측 방법은 다른 예측 모델을 이용하는 것에 비해 크게 두 가지 장점을 가진다. 첫째, 학습된 트리의 해석이 매우 용이하므로 특징 변수들의 중요도, 관계 등을 직감적으로 알 수 있다. 둘째, 학습 자료를 수집하는 과정에서의 특징 변수 미측정, 혹은 측정 오류에 의한 이상자료 (outlier) 자료 문제 등에 견고한 결과를 예측할 수 있다. 최근에는 트리 구조 분류기의 성능을 향상시킬 수 있는 방법이 소개되었으며 그중 본 연구에서는 P&C (perturb and combine) 방법 중 하나인 배깅

책임저자: 이상호 (sangholee@lge.com)
137-900 서울시 서초구 우면동 16번지
LG전자기술원 모바일 멀티미디어 연구소
(전화: 02-526-4113; 팩스: 02-526-4104)

(bootstrap aggregating) 방법과[6] 재구축 트리 (born again tree) 방법을[7] 벡터 회귀 트리에 적용하고 구현한다.

본 연구의 에너지 궤적 생성 방법은 다음과 같다. 우선 음소 수준의 에너지 벡터를 예측한 후 예측된 벡터를 연결한다. 연결된 벡터는 각 음소의 길이에 따라 10 ms씩 다시 표본화되어 최종 에너지 궤적을 구하게 된다. 제안된 방법의 성능을 알아보기 위해 총 39분 분량의 음성 자료를 수집하였고, 객관적 평가 척도에 의해 성능을 검증하였다. 본 논문의 구성은 다음 장에서 벡터 회귀 트리에 대해 설명하고 제안된 방법을 3장에서 실험을 통해 검증한다.

II. 벡터 회귀 트리

일반적으로 알려진 스칼라 회귀 트리는 학습 자료 집합 $\{(x_n, y_n)\}_{n=1}^{N}$으로부터 예측 오류를 최소화하는 방향으로 특징 벡터 x의 공간을 연속적으로 나눈다. 새로운 특징 벡터가 주어졌을 때, 특징 벡터는 비단말 노드에 있는 질문에 의해 트리를 내려오다가 단말 노드에 있는 값을 예측 값 $\hat{y}$으로 결정하게 된다[5]. 본 연구에서는 예측 변수가 벡터인 학습 자료 집합 $\{(x_n, y_n)\}_{n=1}^{N}$으로부터 트리를 학습하고, 벡터를 예측할 수 있는 벡터 회귀 트리를 제안한다.

벡터 회귀 트리를 구현하기 위해 스칼라 회귀 트리 구현시 정의되는 노드 t의 불순도 $r(t)$를 $\frac{1}{N(t)}\sum_{x_n\in t}|y_n-\bar{y}(t)|^2$로 정의한다. $N(t)$는 노드 t에 있는 학습 자료의 개수이며 $\bar{y}(t)$는 노드 내에 있는 목적 벡터들의 평균 벡터이다. 최종적으로 트리 T의 평균 제곱 오류 $R(T)$는 $\frac{1}{N}\sum_{n=1}^{N}|y_n-d(x_n)|^2$로 정의되며 $d(x)$는 입력 특징 벡터 x에 대한 트리 T의 출력 벡터이다.

트리 생성 방법은 트리 확장 단계와 트리 제거 단계로 이루어져 있고, 각각 Chou의 방법과[4] 비용-복합도 제거 방법을[5] 이용하여 구현하였다. 트리 제거 단계에서는 10차 교차 검증 (10-fold cross-validation) 방법이 적용되었고, 최적 트리는 0SE (0 standard error) 방법과 1SE 방법이 모두 적용되었다. 0SE방법은 교차 검증 오류값 $R^{cv}(T)$가 최소가 되는 트리를 찾는 방법이고, 1SE 방법은 $R^{cv}(T)$에 그 트리의 표준 오류값 $SE(R^{cv}(T))$를 더한 값보다 작은 오류를 예측하는 최소 트리를 결정하는 방법이다.

벡터 회귀 트리에 관련된 표현법은 다음과 같다. 트리와 노드는 스칼라 트리와 마찬가지로 각각 T와 t로 표현하고, 트리의 단말 노드 집합은 $\tilde{T}$, 단말 노드 개수는 $|\tilde{T}|$로 표현한다. $R(T)$는 앞에서 정의된 바와 같이 평균 제곱 오류이므로 $\sqrt{R(T)}$는 평균 제곱 오류근 (RMSE)으로 정의된다. 학습 자료에 대한 오류 예측률은 $R^{cv}(T)\pm SE(R^{cv}(T))$로 정의되며 간략하게 $\hat{R}(T)$로 표현한다. 이외에 분산 감소율을 알아보기 위한 상대 평균 제곱 오류 $RE(T)=R(T)/R(\mu)(R(\mu)=[E(y-\mu)^2],\ \mu=E[y])$와, 예측값와 실제값 간의 상관 계수 (correrelation coefficient)을 함께 사용한다. 학습된 트리를 실험 자료에 적용하였을 때는 위의 네가지 표현법에 대해 $R^{ts}(T)$, $\sqrt{R^{ts}(T)}$, $RE^{ts}(T)$, $r^{ts}(T)$로 표현한다.

한편, 본 연구에서는 에너지 궤적 예측률을 높이기 위해 배깅 (bootstrap aggregating) 방법과[6] 재구축 트리 방법을[7] 벡터 회귀 트리에 적용한다. 배깅 방법은 하나의 트리가 아닌 복수 개의 트리를 이용하는 방법이고 재구축 트리 방법은 배깅 트리로부터 다시 하나의 트리를 만드는 방법이다. 이 두 트리들을 위한 표현법은 다음과 같다. 우선 배깅 트리의 경우, 트리들의 집합은 $\{T^b\}$로 표현하고, 트리 집합내 단말 노드 개수 $|\tilde{T}^b|$의 총 합은 $|\tilde{T}^{\{b\}}|$로 표현한다. 배깅 트리에서도 표준 트리의 교차 검증과 유사한 배깅 검증 (out-of-bag) 추정 방법을 통해 오류율의 예측치 $R^{OB}(\{T^b\})$를 구하고[8], $\hat{R}^{OB}(\{T^b\})$는 $R^{OB}(\{T^b\})\pm SE(R^{OB}(\{T^b\}))$로 표현한다. 이외에 표준 트리에서 사용하는 표현법을 배깅 트리에 적용하여 $R^{ts}(\{T^b\})$, $RE^{ts}(\{T^b\})$, $r^{ts}(\{T^b\})$가 정의된다. 재구축 트리의 경우는 오류율 예측치 $\hat{R}(T)$의 계산 방법이 아직 개발되지 않아서 트리의 제거 (pruning) 과정에서 얻어지는 $R^{BA}(T)$를 이용한다. 실험 자료에 관한 오류율 표현법은 표준 트리에서 사용하는 표현법과 동일하다.

III. 에너지 궤적 예측

3.1. 표준 벡터 회귀 트리 이용

제안된 벡터 회귀 트리의 성능을 알아보기 위해 초등학교 교과서, 소설, 논문 등에서 500문장 (4,442어절)을 수집하였고, 여성 아나운서가 방음실에서 발성한 약 39분 분량의 음성 코퍼스를 구축하였다. 문장 코퍼스에 대해 기개발된 문서 분석기를 이용하여 형태소 분석, 발음표기 변환, 구문 분석을 수행하고 분석 오류를 수정하였다. 음성 코퍼스에 대해서는 운율구 경계와 음소 경계를 표시

하였고, 44개의 음소 기호를 이용하여 28,990개의 음소를 얻었다. 본 논문에서는 한국어 발화의 운율 구조는 복수개의 운율구 열로 이루어지고, 하나의 운율구는 운율구 마지막에 놓이는 경계 성조와 나머지 음절에 놓이는 비경계 성조로 이루어져 있다고 가정한다[9]. 그러므로 모든 음절은 자신의 성조를 가지게 되며 수집된 음성에 대해 수동으로 성조를 표기하였다.

음성 코퍼스의 에너지 궤적은 다음과 같이 구해졌다. 우선 매 5 ms의 에너지를 20 ms 구간의 Blackman 윈도우를 사용하여 구하고 에너지를 발화의 최대 에너지에 대해 dB 단위로 표현한다. 그 후, 에너지 열을 5 point median 필터에 통과시키고 낮은 피치의 음성에서 발생할 수 있는 에너지의 급변화를 방지하기 위해 5 point hanning window 필터를 통과시킨다[2]. 최종적으로 각 음소의 길이에 표준화된 음소당 열 개의 에너지 값을 구한다.

벡터 회귀 트리를 학습하기 위해 다음과 같이 총 일곱 개의 카테고리 변수와 세 개의 실변수를 제안한다. 제안된 특징 변수의 첫 글자가 D일 경우는 그 변수가 카테고리 변수임을 뜻하고, C일 경우는 실변수임을 뜻한다.

- Dlph, Dcph, Drph: 이전 음소, 관측 음소, 다음 음소. 이 특징 변수들은 음소 문맥을 반영하기 위해 사용되었다. 관측 음소가 운율구의 처음 혹은 마지막일 경우, 해당 Dlph 혹은 Drph는 NA (Not-Applicable)로 표현된다.
- Dltone, Dctone, Drtone: 음소 문맥과 상응하는 성조 문맥. 성조의 종류가 에너지에 영향을 미칠 것으로 생각되어 사용되었다.
- Dwhineoj: 어절내 음소의 위치. 이 변수는 첫 음절, 중간 음절, 마지막 음절, 세 종류중 하나의 특징 값을 가진다. 만약 하나의 음절로 된 어절일 경우, 마지막 음절로 간주된다.
- Cnsyllphr, Cbsyllphr: 운율구 내 첫 음절 및 마지막 음절로부터의 음절 개수. 이 변수는 에너지가 운율구 내에서 점차 하강한다는 궤적 특성에 기인한다.
- Cnsyllphrr: Cnsyllphr을 운율구 내 음절 개수로 나눈 값.

위 특징 변수들을 이용하여 총 300문장으로 트리를 학습시키고 200 문장에 대해 트리의 성능을 알아보았다. 총 학습 에너지 벡터는 17,618개이었고 실험 에너지 벡터는 11,372개였다. 0SE와 1SE 방법 모두를 적용하였고 우연히 두 트리가 동일하였다. 트리의 성능은 표 1에서 보는 바와 같이 학습 자료에 대한 평균 제곱 오류 (MSE)는 23.73이고 실험 자료의 경우는 22.91이다.

표 1. 에너지 벡터에 대한 표준 벡터 회귀 트리의 성능
Table 1. Performance of the standard vector-regression tree for energy vectors.

Train (N=17618)	Test (N=11372)
$\hat{R}(T)=23.73\pm0.25$ $\lvert\tilde{T}\rvert=105$	$R^{ts}(T)=22.91$ $\sqrt{R^{ts}(T)}=4.78$ $RE^{ts}(T)=0.40$ $r^{ts}(T)=0.77$

표 2. 실험자료의 에너지 궤적에 대한 표준 벡터 회귀 트리의 성능
Table 2. Performance of the vector-regression tree on the energy contours of test utterances.

R^{ts}	$\sqrt{R^{ts}}$	RE^{ts}	r^{ts}
23.85	4.88	0.38	0.78

표에서 보여지는 평균 제곱 오류근 (RMSE)은 에너지 궤적을 음소당 열 개의 에너지 값으로 표현한 후 계산된 것으로, 프레임 당 에너지 예측의 성능을 나타내지 않는다. 그러므로 실제 TTS 시스템에서 이 방법이 사용되었을 때의 성능을 알아보기 위해 200문장의 실험 자료에 대해 10 ms씩 에너지를 생성하고 이를 실제 에너지 궤적과 비교하였다. 에너지 벡터로부터 문장의 에너지 궤적을 구하는 방법은 우선, 연접하는 두 음소의 에너지 벡터 e_n과 e_{n+1}이 주어졌을 때 연접 위치의 에너지 값을 $(e_{n,10}+e_{n+1,1})/2$로 설정한다. 그 다음 에너지 벡터들을 음소 길이에 맞추고 선형 보간법에 의해 10 ms 씩 에너지 값을 계산한 후 3 point hanning window 필터를 통과시킨다. 실험 자료의 총 78,161개의 프레임에 대해 성능을 알아본 결과 표 2의 결과를 얻었다.

본 연구에서 제안하는 벡터 회귀 트리가 에너지 궤적 예측 문제에 효과적인지를 알아보기 위해 기존 연구자들이 보고한 결과와 비교하였다. Bagshaw는 음소 수준에서 6.8 dB의 평균 제곱 오류근 (RMSE)을 얻었으며[2], 이는 표 2에 보여진 프레임 수준에서의 4.88 dB보다 더 높은 값이다. Ross와 Ostendorf는 48분 분량의 음성에 다이나미컬 시스템을 학습시키고 11분 분량의 음성에 실험한 결과, 프레임 수준에서 3.48 평균 제곱 오류근을 얻었다. 하지만 그들의 실험 자료 표준 편차는 4.88 dB이었으며 결과적으로 상대 평균 제곱 오류는 $(3.48/4.88)^2 = 0.508$이 되며, 이 값은 본 실험에서 얻은 0.38보다 높은

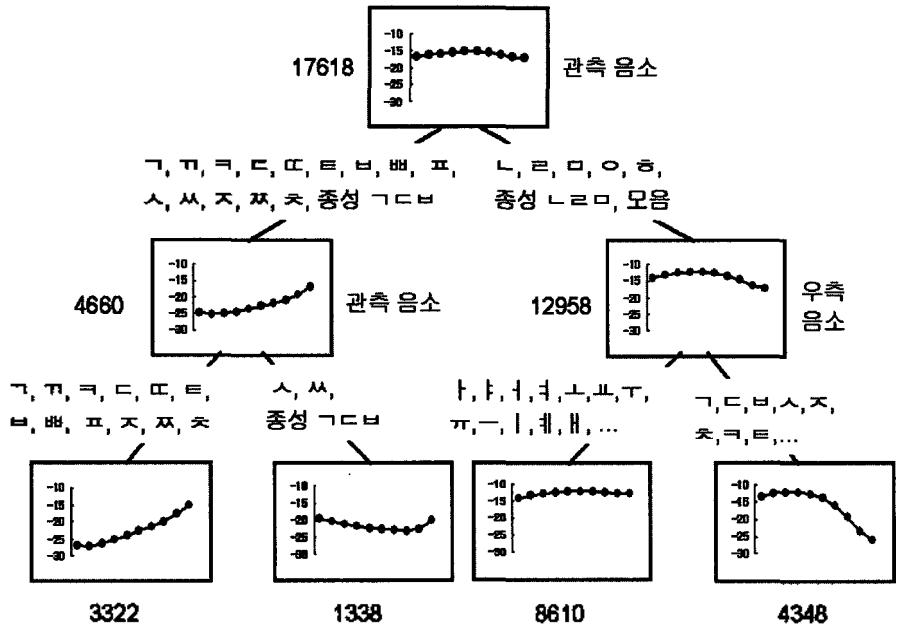


그림 1. 에너지 벡터 예측을 위한 표준 백터 회귀 트리
Fig. 1. Standard vector-regression tree for energy vector prediction.

깊이다.

구해진 벡터 회귀 트리에서 루트 노드로부터 손자 노드까지의 구성은 그림 1과 같다. 그림에서 노드 안의 에너지 궤적은 그 노드에 도달한 에너지 벡터들의 평균 벡터를 보인 것이다. 루트 노드의 질문은 관측 음소에 관한 것으로 관측 음소를 무성음과 유성음으로 나누는 것을 볼 수 있다. 좌측 노드는 초성 자음과 종성 자음이 있으며 우측 노드는 모음과 유성 자음이 있는 것을 알 수 있다. 그림에서는 루트로부터 몇 개의 노드들만 나타나 있어 사용된 특징 변수들이 음소 정보뿐이지만, 최종 트리는 본 연구에서 제안한 모든 특징 변수를 이용한다.

3 2. 향상된 백터 회귀 트리 이용

앞에서 구한 표준 벡터 회귀 트리의 성능을 높이고자 본 절에서는 배깅 방법과 재구축 트리 방법을 이용하여 트리를 학습시킨다. 배깅 트리는 학습자료로부터 열 개의 표본 추출된 학습 자료들로부터 만들어졌으며, 재구축 트리는 20% 혼합 (smearing) 방법에 의해 만들어졌다. 두 트리들의 성능은 표 3에 나타나 있다. 표에서 보는 바와 같이 실험 자료에 대해 상관 계수가 기존의 0.77에서 0.79로 올라가는 것을 볼 수 있다. 표 4에서는 실험 자료의 총 78,161개의 프레임에 대한 성능을 보이고 있으며 재구축 트리의 경우, 기존의 0.78에서 0.803으로 올라가는 것을 확인할 수 있다.

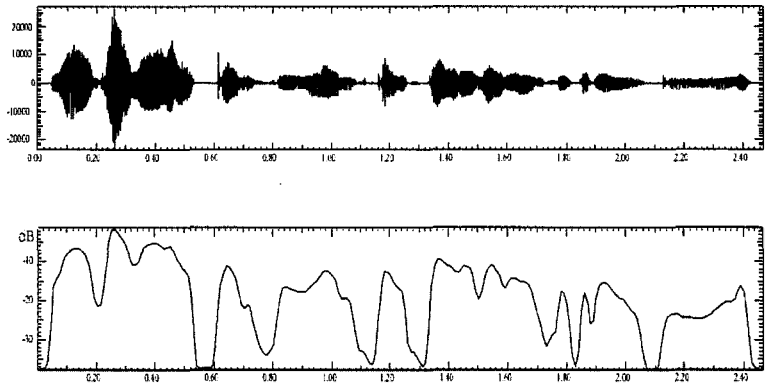
최종적으로 만들어진 재구축 트리의 성능을 청각적으로 평가하기 위해서, 임의의 실험자료 음성 신호에 대해 그 신호의 에너지 궤적을 트리가 예측한 궤적으로 바꾸어

표 3. 에너지 백터에 대한 향상된 백터 회귀 트리의 성능
Table 3. Performance of the improved vector-regression trees for energy vectors.

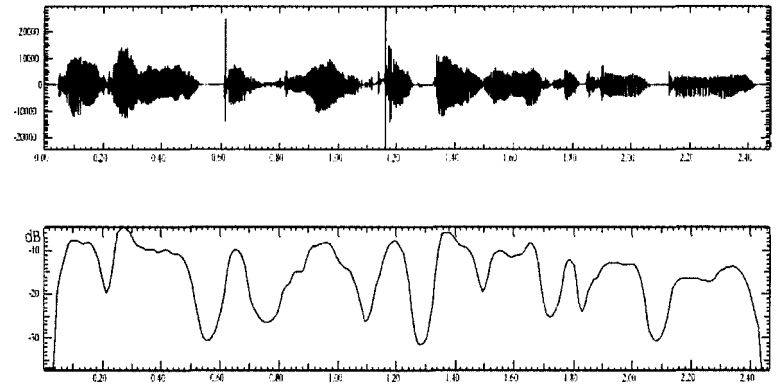
	Train (N=17618)	Test (N=11372)
BAG	$\hat{R}^{OB}(\{T^b\}) = 20.41 \pm 0.22$ $\lvert \hat{T}^{(b)} \rvert = 67513$	$R^{ts}(\{T^b\}) = 20.68$ $\sqrt{R^{ts}(\{T^b\})} = 4.548$ $RE^{ts}(\{T^b\}) = 0.367$ $r^{ts}(\{T^b\}) = 0.799$
BA	$R^{BA}(T) = 18.96$ $\lvert T \rvert = 599$	$R^{ts}(T) = 21.17$ $\sqrt{R^{ts}(T)} = 4.601$ $RE^{ts}(T) = 0.376$ $r^{ts}(T) = 0.791$

표 4. 실험자료의 에너지 궤적에 대한 향상된 벡터 회귀 트리의 성능
Table 4. Performance of the improved vector-regression trees on the energy contours of test utterances.

	BAG	BA
R^{ts}	21.55	22.31
$\sqrt{R^{ts}}$	4.643	4.723
RE^{ts}	0.810	0.803
r^{ts}	0.345	0.357



(a) 원음 및 에너지 궤적
(a) Original speech and its energy contour



(b) 변형된 합성음 및 예측된 에너지 궤적
(b) Transformed speech and predicted energy contour

그림 2. 원음과 예측된 에너지 궤적을 가지는 합성음
Fig. 2. Original speech and the transformed speech with the predicted energy contour.

합성하였다. 원음과 변형된 음성 신호는 청각적으로 구분할 수 없었으며, 그림 2에 두 개의 음성 파형과 각각의 에너지 궤적을 보여주고 있다. 그림 2는 "물건을 아껴쓰면 어떤 점에서 좋을까요?"를 발화한 음성 파형들이다. 그림에서 보는 바와 같이 두 개의 에너지 궤적이 매우 비슷한 것을 알 수 있다.

IV. 결 론

본 논문에서는 한국어 에너지 궤적 예측을 위한 벡터 회귀 트리에 대해 논하였다. 벡터 회귀 트리의 입력으로는 음소 문맥, 성조 문맥, 운율구내 관측 음소의 위치 등을 이용하였고, 음소당 정규화된 열 개의 에너지 값을 예측하여 전체 에너지 궤적을 구할 수 있었다. 실험은 300문장의 학습 자료와 200문장의 실험자료를 통해 이루어졌으며, 객관적 평가 척도에서 벡터 회귀 트리는 기존의 방법들보다 우수하였으며 주관적 평가에서도 청각적으로 원음과의 차이를 느낄 수 없었다. 최종적으로 사용된 재구축 트리는 표준 트리보다 성능이 우수하지만 트리가 크다는 단점이 있다. 앞으로 트리 크기 감소에 대한 연구가 필요하다.

참 고 문 헌

1. J. C. Lee, D. G. Kang, S. H. Kim and K. M. Sung, "Energy contour generation for a sentence using a neural network learning method," *Proc. Int. Conf. Spoken Language Processing*, 1991-1994, 1998.
2. P. C. Bagshaw, "Unsupervised training of phone duration and energy models for text-to-speech synthesis," *Proc. Int. Conf. Spoken Language Processing*, 17-20, 1998.
3. K. Ross and M. Ostendorf, "A dynamical system model for generating fundamental frequency for speech synthesis," *IEEE Trans. Speech Audio Processing*, **7**, 295-309, May 1999.

4 P. A. Chou, "Optimal partitioning for classification and regression trees," *IEEE Trans. Pattern Anal. Machine Intell.*, **13**, 340-354, Apr. 1991.
5 L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, ser. Wadsworth Statistics/ Probability Series Belmont, CA, 1984.
6. L. Breiman, "Bagging Predictors," *Machine Learning*, 24, 123-140, 1996.
7 L. Breiman and N. Shang, Born Again Trees, ftp://ftp.stat.berkeley.edu/pub/users/breiman/BAtrees.ps, 1996.
8. L. Breiman, Out-of-Bag Estimation, ftp://ftp.stat.berkeley.edu/pub/users /breiman/OOBestImation.ps.Z, 1996.
9. 이상호, 오영환, "한국어 억양의 트리 기반 모델링," 한국음향학회지, **19** (2), 19-32, 2000.

저자 약력

● 이 상 호 (Sangho Lee)

1993년 2월: 동국대학교 전자계산학과 (학사)
1995년 2월: 한국과학기술원 전자전산학과 (석사)
2000년 2월: 한국과학기술원 전자전산학과 (박사)
2000년 3월~현재: LG전자기술원 모바일 멀티미디어연구소 선임연구원
※ 주관심분야: 음성인식, 음성합성, 자연언어처리, 패턴인식

● 오 영 환 (Yung-Hwan Oh)

1972년: 서울대학교 공과대학 (학사)
1974년: 서울대학교 교육대학원 (석사)
1980년: Tokyo Institute of Technology 정보공학전공 (박사)
1981년~1985년: 충북대학교 컴퓨터 공학과 조교수
1983년~1984년: University of California (Davis) 연구교수
1995년~1996년: Carnegie-Mellon University 연구교수
1985년~현재: 한국과학기술원 전자전산학과 전산학 전공 교수
※ 주관심분야: 음성인식, 음성합성, 음성코딩, 화자인식, 대화관리, 신경회로망, 전문가시스템