A Study on Lattice Rescoring with Knowledge Scores for Automatic Speech Recognition

Authors: *Sabato Marco Siniscalchi, Jinyu Li, and Chin-Hui Lee*

Professor:陳嘉平
Reporter: 吳柏鋒

Outline

- Introduction
- Knowledge-based scores
  - Computing LLR scores
  - Phone-Level scores
- Lattice rescoring
- Experiment

Introduction

- Frame-based log likelihood ratio is adopted as a score measure of the goodness-of-fit between a speech segment and the knowledge sources.

- Knowledge scores obtained from 15 attribute detectors for place and manner of articulation.

  that were realized with feed forward artificial neural networks (ANNs).

Introduction

- Propose knowledge scores based upon log likelihood ratio (*LLR*) that can be computed at a segment, or frame level.

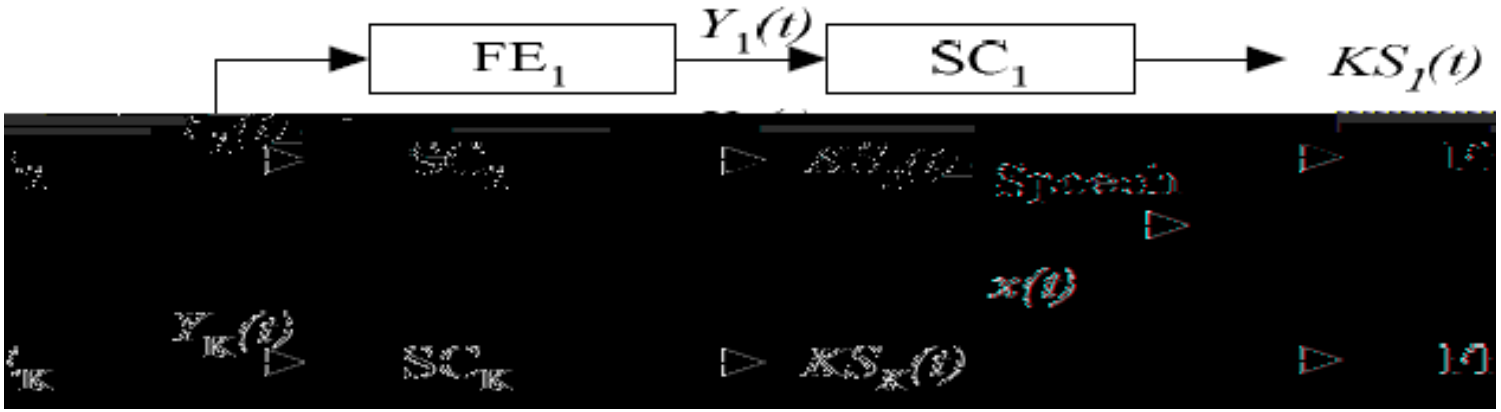
- Phone-level scores are obtained as a nonlinear mapping of *LLR* scores into the phone space. These phone scores are used to rescore lattices of alternative hypotheses.

Knowledge-based scores

- We use articulatory information as knowledge source that features are related to human speech production

- ANNs are often adopted to map MFCCs into articulatory information, because their output can be the a posteriori probability of observing an articulatory attribute for the given input.

Knowledge-based scores



A_i, FE_i stands for a feature extraction module that converts a speech signal $x(t)$ into a sequence of speech parameter vectors, Y_i. SC_i is an attribute scoring module that computes knowledge score, KS_i.

Computing LLR scores

- LLR can be computed at a segment level, $LLR^{(s)}$, or at a frame level, $LLR^{(f)}$.
  - **$LLR^{(s)}$**

Computing LLR scores

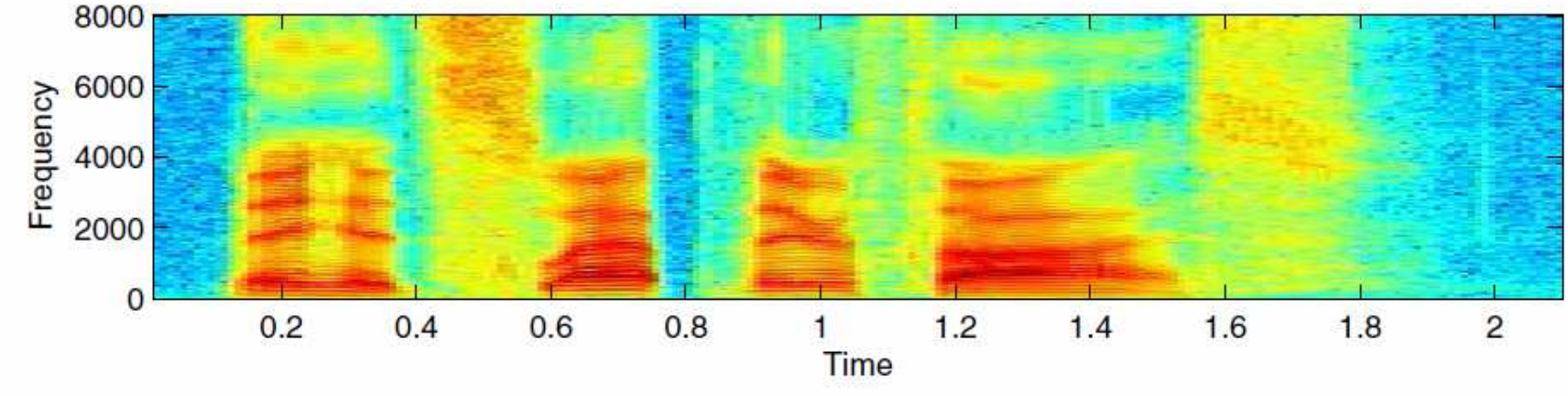
- The $LLR^{(f)}$ can be defined as:

$$LLR_i^{(f)}(o_t) = log\frac{P(o_t|\lambda_i)}{P(o_t|\lambda_i^a)}$$

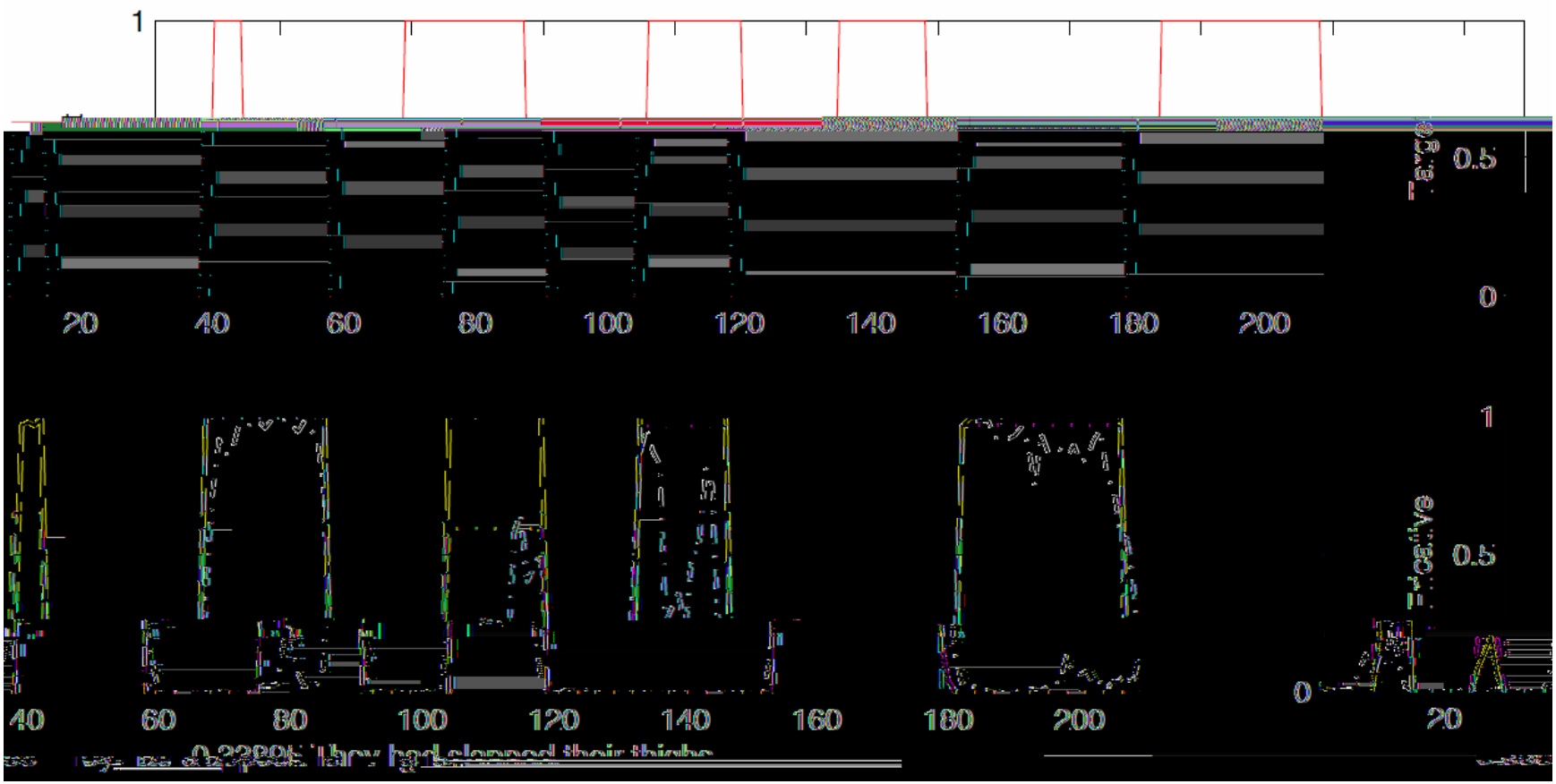
generated by an observation vector o_t in state i at time t and λ_i is the target HMM model of the articulatory class ending in state i, and λ_i^a is its corresponding competing model.

Computing LLR scores

- compares segment and frame detectors for the fricative manner. In order to report all the scores to the same range of values, we apply a sigmoid limiter to the *LLR* scores.



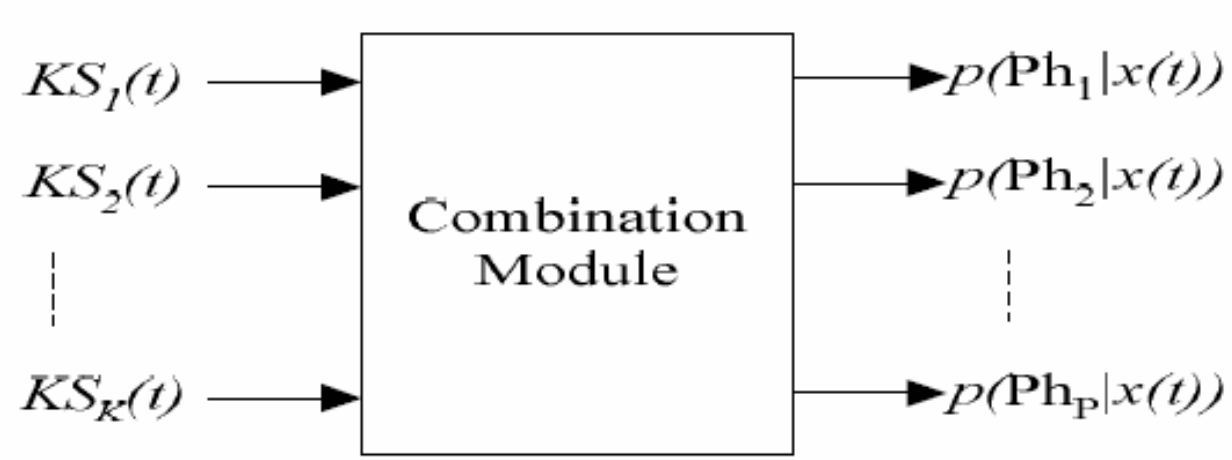
Computing LLR scores



The figure is detection curves for the fricative attribute.
Represented ANN-based, LLR $^{(f)}$, LLR $^{(s)}$ scores

Phone-Level scores

- We use a non-linear function realized by a feed forward ANN with one hidden layer with 100 hidden nodes.



Given the m-th frame and the n-th candidate string, we obtain the

Lattice rescoring

- We denoted the rescoring log likelihood value as:

$$S_n = w_{ps} * PS_n + w_l * L_n$$

where L_n is the log likelihood of the *n-th* arc; PS_n is a linear combination of $PS_{n,m}$ for each arc, with $PS_{n,m}$ being a non-linear transformation of the score of the *m-th* frame for the *n-th* arc and set the w_{ps} and w_l are set to be equal.

Experiment

- Use the TIMIT database.
- HMM based detectors for 15 speech attribute, namely fricative, vowel, stop, nasal, approximant, low, mid, high, labial, coronal, dental, velar, retroflex, glottal, and silence.
- Each HMM has 3 states with 32 Gaussian mixture components per state.

Experiment

- HTK is used to build the context-Independent (CI) and context-dependent(CD) baseline phone recognition systems.

Table 2: *Phone lattice scoring performance*

Phone error rate	CI Phone	CD Phone
Baseline	40.52%	36.13%
Rescore	35.16%	33.42%