A NEW COMBINED MODEL OF STATICS-DYNAMICS OF SPEECH

Zhijian Ou, Zuoying Wang

Department of Electronic Engineering, Tsinghua Univ., Beijing, China

❑ Abstract

√ Linear prediction (LP) HMM does not make the independent and identical distribution (IID) assumption in traditional HMM; however it often produces unsatisfactory results.

√ In this paper, a new combined model of statics-dynamics of speech is proposed. It works with **LPHMM** as the dynamic part and traditional **IID-based HMM** as the static part.

❑ Linear Prediction HMM

- Generally suppose the D-dimension observation o_t within a state s is described as

$$o_t = \sum_{i=1}^{m} \beta_i^s o_{t+l_i} + \mu_s + \nu_t$$

l_i : the "offset" associated with the i^{th} predictor;

$\beta_i^s \in R^{D\times D}$: the i^{th} prediction matrix;

$\mu_s \in R^D$: accounts for a non-zero mean of the observations;

$\nu_t \sim \mathrm{N}(0,\Sigma_s)$: Gaussian noise (un-correlated between frames).

- For state s, the output probability density function (pdf) of observation o_t then becomes correlated, conditional on its context $\{o_{t+l_1}, \cdots, o_{t+l_m}\}$:

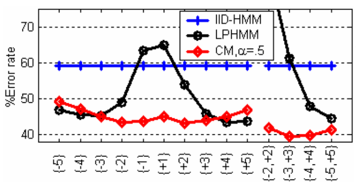
$$\tilde{b}_s(o_t) \stackrel{\Delta}{=} p(o_t \mid o_{t+l_i}, i=1,\cdots,m,s)$$

$$= \frac{1}{(2\pi)^{D/2}|\Sigma_s|^{1/2}} \exp\left\{-\frac{1}{2}\left(w_t^s-\mu_s\right)^T \Sigma_s^{-1}\left(w_t^s-\mu_s\right)\right\},$$

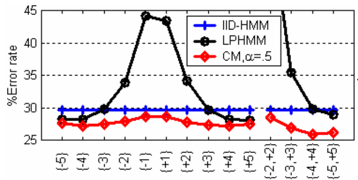
where $w_t^s = o_t - \sum_{i=1}^{m} \beta_i^s o_{t+l_i}$.

❑ EXPERIMENTAL RESULTS

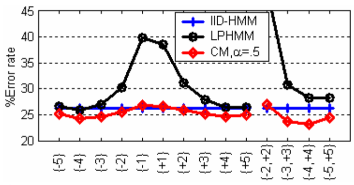
Chinese speaker-independent continuous speech recognition:



←15 dim: basic feature, including 14 MFCCs and normalized log-energy.



←30 dim: $+\Delta$



←45 dim: $+\Delta\Delta$

The combined model (CM) achieved better performance than both IID-HMM and LPHMM.

Under 45 dim, the CM using {-4,+4} achieved 11.4% relative error rate reduction from IID-HMM. (from 26.30% to 23.30%)

↑ Average error rates for various models, each with specific feature dimension, *model type* and $\{l_1,\cdots,l_m\}$.

❑ A Combined Model

➢ Analysis

⅄ For parameter estimation, LPHMM is to minimize the determinant of the sample covariance of $o_t - \sum_{i=1}^{m} \beta_i^s o_{t+l_i}$, i.e., to find such β_i^s's that o_t is most compactly distributed conditional on its context (or say, around the value of $\sum_{i=1}^{m} \beta_i^s o_{t+l_i}$.) In this way, the *dynamics* of outputs of state s is well captured in LPHMM embodied by the correlated output pdf $\tilde{b}_s(o_t)$.

⅄ On the other hand, traditional IID-based HMM is still effective in practical speech recognition, maybe due to its good ability at modeling the *statics* of speech. All the observations in each state are well statically (unconditionally) distributed in a cluster represented by the mean of the standard output pdf $b_s(o_t)$, regardless of any nearby observations.

⅄ The weak points are that, to decide which state the feature o_t most probably comes from, the matching score computed by $\tilde{b}_s(o_t)$ alone is insufficient, if the matching score by $b_s(o_t)$ is not taken into account, and vice versa.

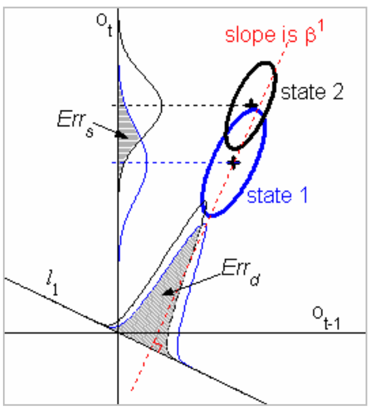
➢ Formulation

Neither LPHMM nor IID-based HMM alone is sufficient. It is beneficial to utilize the complementary modeling powers on statics and dynamics of speech of these two kinds of HMMs to yield a combined model. The new "combined output pdf" is defined as

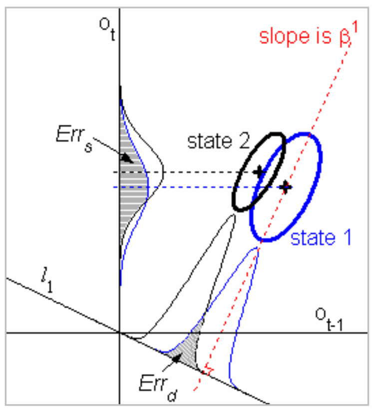
$$\tilde{\tilde{b}}_s(o_t) = b_s(o_t)^{1-\alpha} \cdot \tilde{b}_s(o_t)^{\alpha}.$$

➢ Illustration

- Here o_t is regarded as one-dimensional and $m=1, l_1=-1$. Each ellipse is the contour line of $p(o_t, o_{t-1} \mid s)$, conceptually characterizing the output features of each state $s=1,2$.
- The Gaussian pdf curves along the o_t axis and the sloping line l_1 respectively represent $\{b_1(o_t), b_2(o_t)\}$, and $\{\tilde{b}_1(o_t), \tilde{b}_2(o_t)\}$ that are put together along l_1 for clear view.
- The overlapping area of two pdf curves gives the classification error.
- Using $b_s(o_t)$ and $\tilde{b}_s(o_t)$ yields Err_s and Err_d respectively.



↑ An illustration of how LPHMM fails to discriminate between two states, where $Err_s < Err_d$.



↑ An illustration of how IID-based HMM fails to discriminate between two states, where $Err_s > Err_d$.