Improving Hidden Markov Model Performance in Phoneme Classification by Fuzzy Smoothing

[1]FARBOD HOSSEYNDOOST, [2]MOHAMMAD TESHNEHLAB

[1]Department of Artificial Intelligence,
Science and Research Campus, Azad University

[2]Department of Electrical Engineering,
K.N.Toosi University of Technology

Abstract: In this paper, two kinds of uncertainties, in speech production and recognition, are introduced. It is shown that one class of these uncertainties can be best understood by the notion of probability while the other could be described as fuzziness. Based on the given concepts, a new method of fuzzy smoothing is proposed. The goal of this method is to make transition from one state to another gradual, so that contribution of each observation to the emission probability becomes fuzzy. In addition, an innovative implementation method is suggested. It is shown that adding a length normalized time dimension to the feature space can serve as fuzzy smoothing. Error rates in phoneme classification are compared on TIMIT database. Results show, a significant improvement in classification rate over the original HMM, while imposing no computational intricacy.

Key-Words: Hidden Markov Model, Speech Recognition, Fuzzy Logic, Smoothing, Phoneme Classification

1 Introduction

Hidden Markov Model (HMM) has shown successful in many applications including speech recognition and speaker identification for quite a long time. Nevertheless, it has never been claimed perfect for speech modeling. There are some drawbacks in HMM structure as well as its training methods. For one thing, it is unable to compare two non-equal-length observation sequences, since probability inherently lies between 0 and 1, and probability of observations given the model declines over the time. This problem reveals itself, in segmentation and verification while has no effect on recognition and classification where one string, is scored by many HMMs. Second, using original Baum-Welch re-estimation method, one HMM's Gaussians, are trained, without considering distribution functions of rival HMMs. In recognition, success criterion is not increasing score of observations given the model, but maximizing score distance between rival models. Finally (in the interests of brevity) the transition between the states, in HMM, although hidden, is sudden. As we will describe, it exhibits itself, incompatible with the utterance nature.

The first problem, stated above, is not recognized as crucial and not explicitly discussed. The second problem is addressed in many papers, and the solution to the problem has formed a new class of models under the title of discriminative HMMs. Neural Networks [7] and Support Vector Machines [4] are used, successfully in estimating emission probability in HMM, by classifying rival patterns and using new criteria as it is done in MMI[2].

The third problem is also addressed in some ways, under the name of smoothing or soft segmentation, but not in the context of HMM[3]. It seems that in HMM, the general assumption is that emission probability, and unclarity of state sequence, works for modeling the gradual transition from one state to another. We try to show that there are two types of uncertainty in speech production and they need to be handled by different means. The idea of smoothing, could be, best described by fuzzy concepts, while the traditional probabilistic framework of HMM is untouched.

HMM equations are reformulated by fuzzy concepts, replacing the notion of probability by possibility in fuzzy logic realm and fuzzy HMMs are proposed in some papers[5,6]. Our approach is, quite different. We give a new description of the problem and show that the two types of uncertainties are different. Additionally, we suggest an innovative technique to implement the new model, with a little change in original HMM structure and re-estimation formula. We try to reconcile smoothing method, with hidden Markov model and test the new model on TIMIT database. It is also notable that this

technique is length independent, and can alleviate the wrong duration modeling of HMM.

2 Fuzzy Hidden Markov Model

Pronouncing a phoneme, there is delay from the moment we intend to produce a sound, to the time we do it. It is different, in essence, from the concept of many states in HMM, which each state, is meant to correspond to a certain position of articulation. For example, we know from IPA chart that, pronouncing phoneme, *ow,* there is a transition from one manner and position of articulation to another. We can stay, in one state of articulation, for as much as needed (as singers do) but there is again a transition interval, from that state to the next.

To clarify the problem, suppose that we want to work out the probability of an office worker being on the first floor of a three-story building during the day. We define the probability of being on the first floor, as the time spent on the first floor over the time spent on any floor. Also, imagine that our observations are scarce and most of the time the worker, is moving from one floor to another, so we often, see him in the staircase. While being on any floor is a probabilistic variable, it is a fuzzy one, as well. We aim at the probability of being at the first floor, but our observations that indicating being on the first floor, are mostly, fuzzily true. The question is that, what the probability of being on the first floor is, given that we know, the probability of being in the middle of the staircase. In other words, we want to know, how much the time, spent in the middle of the staircase, can contribute to the probability of being on the first floor.

2.1 Smoothing by Fuzzy Probability

In original HMM formulation, the goal of EM method is to increase the probability of all observation sequences given the model, where probability of an observation sequence given the model and one explicit path on HMM is:

$$P(O^j \mid M, Q) = \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \qquad (1)$$

O^j is a sequence of observations, o_i , with the length T. in other words, $O^j = \{o_1, ..., o_T\}$.

a is transition probability and a_{0i} is the prior probability of state i (probability of starting from state i) and q_0=0. Q is the path or sequence of states on HMM starting from q_1 and ending in q_T.

Probability of observations given the model, $P(O|M)$ is sum of this probability over all possible sequences of states(Q).

In Expectation-Maximization re-estimation, the goal is to maximize the value of:

$$P(Oself \mid M) = \prod_{j=1}^{n} P(O^j \mid M) \qquad (2)$$

where $Oself$ is the set of independent same-class observations, for example, all instances of a phoneme. O^j is an observation sequence, which may be, feature frames of one instance of pronounced phonemes.

In traditional HMM the probability of a frame of observation given a state is defined as emission probability:

$$b_j(o) = P(O = o \mid s = s_j) \qquad (3)$$

We suggest that this concept is as fuzzy as probabilistic. There is no need to replace one concept for another. Two types of observation could be imagined of: Intended and Uttered. Intended observation is that, shapes the base characteristics of the state; i.e. the perfect representation of the state. It is, at the same time, probabilistic and can have a distribution function. Uttered observation, relates to the Intended observation with a fuzzy membership function. The goal of re-estimation should be raising the probability of Intended observations given the model. In this view, the observations, with lower fuzziness, contribute less to form the Intended observation distribution function of states. We can ask what the probability of Intended observation is, given that we know the probability of Uttered observation, and fuzziness of Uttered observation. It is similar to asking: We know, how much, 'being in the middle of the staircase', is, 'being on the first floor' and we know, what the probability or time of being in the middle of the staircase is, during the day and we want to know how much this, can participate in the probability of being on the first floor. There are many ways to define fuzzy probability. We propose two ways and believe that both are reasonable. As one definition we can write:

$$P(Intended) = P(Uttered).F_{Int}(Uttered) \qquad (4)$$

Where, $F_{Int}(Uttered)$ is the fuzziness of Uttered, based on Intended or Intended-ness of Uttered.

Another way to define fuzzy probability is:

$$P(Intended) = P(Uttered)^{1/F\,\mathrm{int}(Uttered)} \qquad (5)$$

Note that, probability is a value between zero and one, so when the fuzziness goes down, the probability, decreases. Both definitions are defendable. Using (5) we get closer to the smoothing and soft segmentation [3] where the membership appears in power. We opt to use equation (4) because, as we show later, there is a straightforward computation method for that, based on HMM equations.

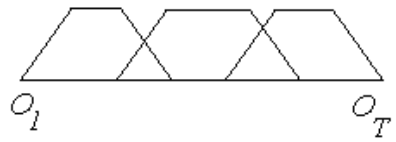
We try to maximize the probability of Intended observations given the model, having the probability of Uttered observation given the model. In recognition, with the same process we work out the probability of Intended observations, given different models, and select the best one. To maximize:

$$P(O_{Intended} \mid M, Q) = \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_{t\_Uttered}) F_{q_t,t}(o_{t\_Uttered}) \quad (6)$$

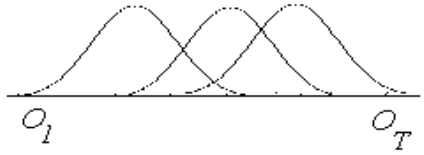
$o_{i\_Uttered}$ is uttered (seen) observation at time *t*. We should re-estimate both membership functions (F), and model parameters using EM. As it could be noted, we have defined membership function of each state, as a function of time.

2.2 Membership Functions

The membership functions(MFs) may be chosen from any form.



(a) Trapezoidal MFs



(b) Gaussian MFs

Fig 1: Membership functions for three states

Fuzziness of an Uttered observation can be defined as its distance from Intended observation or it could be defined over the time (as we choose). The center and width or variance of each MF should be re-estimated in EM iterations.

One option is to keep the center and width of MFs, constant. It forces, the state centers to move to that constant position, through different EM iterations. Another is to set centers of MF for states *q*, as weighted mean of time of observations belonging to state q:

$$Center(F_q) = 1/T \sum_{t=1}^{T} t.p(s_t = q \mid O, M) \quad (7)$$

for Gaussian membership functions, Variance, is estimated by the following equation:

$$Var(F_q) = 1/T \sum_{t=1}^{T} (t - center)^2 .p(s_t = q \mid O, M) \quad (8)$$

2.3 Training Formula

There is just one deviation from original HMM formulations, in our proposed, smoothing method. We avoid rewriting all forward-backward and EM formula, for briefness. The only change to all the original HMM formula in [1] is that $b_{q_i}(o_i)$ will be replaced by $b_{q_i}(o_{i\_Uttered}) F_{q_i,i}(o_{i\_Uttered})$. In EM iterations, first, model parameters (Priors, Transition Probabilities and Emission Probabilities) are estimated and based on these values, using (7) and (8), membership functions are approximated

2.4 Realization Method

We noted that there is a shorter route to estimation of fuzzy membership functions and model parameters, all together. In order to keep away from unnecessary complication, first we assume that each state has only one Gaussian function. We add one dimension to the observation space. This new dimension is normalized time, from *1/T* to *1* added to each observation sequence, $O^j = \{o_1, ..., o_T\}$. We name this new sequence *X* where *X=[O^j;Tm]* and *Tm* denotes a row containing values from *1/T* to *1*. Each column contains a frame of observation and O^j is an observation sequence. ';' symbolizes next row.

Since diagonal Gaussians are separable, for state q, we can write:

$$N(X, \mu_{Xq}, \delta_{Xq}) = N(O^j, \mu_{oq}, \delta_{oq}).N(t, \mu_{tq}, \delta_{tq}) \quad (9)$$

It is clear that the first term on the right side of (9) is $b_{q_i}(o_{i\_Uttered})$ and the second one, is MF for state q. When more than one Gaussian functions are used to estimate distribution probability of each state, we can think of, second term as fuzziness of observation according to state *q* and mixture *m*.

3 Phoneme Classification Results

In this section, we report the results of phoneme recognition, using above mentioned, training method. We use all phonemes of male speakers of first dialect in TIMIT database, Train set, and test our HMMs on the Test set of the same dialect, male speakers. Traditional HMM results are mentioned along with our suggested HMM's results. We had observed that, the big number of some phonemes, with high recognition rate, affects the error rate significantly. To avoid this, as in our previous works, we test min(50, number of phoneme instances) phonemes in the test database. The total number of phonemes, tested, is 1626 in all the experiments. Phoneme classification error is wrong detection of phoneme class. 60 phonemes are classified into 39 phoneme classes. MFCC features are used, with delta, delta delta coefficients and log energy (40 features). A pre-emphasis filter is used as usual before feature extraction ($1\text{-}0.95z^{-1}$). Number of iterations in training all HMMs is 12. All HMMs are left to right with prior probability of first state, 1, and the rest 0.

By adding more than one 'normalized time row' to the end of our observation sequence, we are weighting results from fuzzy part over probability part. In table 1, *tr*, denotes number of time rows added to the observation sequence. *Q* is the number of states and *M*, shows number of mixtures. T-HMM marks, Traditional HMM and F-HMM is our proposed fuzzy HMM. *notrans* means, taking out transition probability from HMM equations and making transition to self or another state, equally likely.

Experiment	Q	M	Errors	%PER
T-HMM	3	5	570	35.06
F-HMM *tr=1*	3	5	554	34.07
F-HMM *tr=4*	3	5	507	31.18
F-HMM *tr=7*	3	5	496	30.50
F-HMM *tr=8*	3	5	497	30.57
F-HMM *tr=12*	3	5	506	31.12
F-HMM *tr=24*	3	5	527	32.41
T-HMM notrans	3	5	595	36.59
F-HMM notrans *tr=8*	3	5	493	30.32
T-HMM	1	5	653	40.16
F-HMM *tr=8*	3	1	648	39.85
F-HMM *tr=8*	1	5	597	36.72

Table 1 : Phoneme Classification Error Rate

It is noteworthy that, in this method, transition probabilities do not play an important role and could be eliminated (Table 1, compare Rows 6, 10).

Removing transition probabilities, gives one advantage, and that is making length normalization easier by dividing the results by the length of observation sequence. It would be helpful in the applications, where two non-equal length observations should be compared with one HMM. Two of those applications are automatic phonetic transcription and verification. Investigating and comparing the effect of such normalizations needs separate research and discussion.

4 Conclusion

In this paper, we formulated a smoothing technique with fuzzy ideas. Soft Segmentation and smoothing techniques, had been examined before, with some variations such as keeping the smoothing function center constant or moving, but our approach was inherently different. We put to use fuzzy notion, in addition to probability to explain the same fact. Instead of using smoothing factor, in power, we practiced this factor as a multiplying coefficient, and showed briefly that adding a normalized time feature to the feature space, can serve as a fuzzy smoothing method. We tested this idea on TIMIT database and showed that the results are significantly improved, compared to those of traditional HMM.

References:

[1] L.R.Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceeding of IEEE, Vol 77, No 2, pp 257-286,1989

[2] Wu Chou, "Minimum Classification Error(MCE) Approach in Pattern Recognition", Pattern Recognition for Speech and Language Processing, CRC Press, 2003, pp 1-40

[3] Farbod Razzazi, Abolghasem Sayadiyan, "Phoneme Recognition System Based on Soft Segment Modeling", ISPC 2003, Dallas, USA

[4] Steven E. Golowich, Don X.Sun, "A Support Vector/Hidden Markov Model Approach to Phoneme Recognition", Bell Labs, 1998

[5] Adrian David Cheok, Sylvain Chevalier, "Use of a Novel Generalized Fuzzy Hidden Markov Model for Speech Recognition"

[6] Jia Zeng, Zhi-Qiang Liu, "Type-2 Fuzzy Hidden Markov Models to Phoneme Recognition", Pattern Recognition, 17th International Conference on (ICPR'04) Vol. 1

[7] Shigeru Katagiri, "Speech Recognition using Neural Networks", Pattern Recognition for Speech and Language Processing, CRC Press, 2003, pp 115-146