SPEECH UNITS AUTOMATICALLY GENERATED BY ERGODIC HIDDEN MARKOV MODEL

Ing. Igor Szöke, Doctoral Degree Programme (1)
Dept. of Computer Graphics and Multimedia, FIT, BUT
E-mail: szoke@fit.vutbr.cz

Supervised by: Dr. Jan Černocký

ABSTRACT

Units automatically generated are used in VLBR (Very Low BitRate) coding. One approach to generate ALISP[1] units was presented in [4]. Temporal decomposition and vector quantization were used there. Our goal is to propose better approach. We use EHMM[2] for ALISP units determination. A brief description of the model, its initialization and training is discussed in this paper. Experimental results are discussed in the conclusion.

1 INTRODUCTION

ALISP units are needed in VLBR coding. The coder works on recognition-synthesis principle. To make the coder language independent, phonemes or other related units can not be used (because they are language dependent). Another potential application of ALISP units is in language identification.

2 ERGODIC HIDDEN MARKOV MODEL

Linear models with three emitting states are usually used in speech processing. Such model is shown in figure 1a. Each model represents a phoneme or a diphone in recognition. Model which has the best likelihood[3] for a part of speech signal is selected and we can say, that this part of speech signal contains the phoneme (diphone) which the model represents. Before we can recognize, we must train models. Models are usually trained with labelled speech data.

Ergodic model in comparison with standard models is fully connected (figure 1b). We use only one ergodic model for recognition. The recognizer outputs a stream of states. Each state represents a part of speech (automatically generated unit).

[1] ALISP - Automatic Language Independent Speech Processing

[2] EHMM - Ergodic Hidden Markov Model

[3] Likelihood means probability that a model emits given part of speech signal.

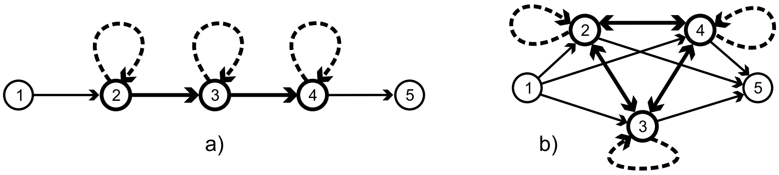

Figure 1: Hidden Markov models: a) standard, b) ergodic

2.1 EHMM PROTOTYPE

Model prototype contains an information about data format and model. Each state has one gaussian mixture. More mixtures in one state are equivalent to one mixture and more states of EHMM. Next part of the prototype describes mixture (gaussian) variances and means. The last part defines transitions costs. It is a matrix of probabilities of transition between states. All of these values (means, variances, transitions probabilities) are unknown. We set them in initialization phase.

3 INITIALIZATION

The initialization is important for EHMM's successful usage. Correctly trained EHMM should assign only one state to a set of acoustically similar segments. The following approaches were used for initialization of gaussian mixtures variances and means:

Constant/Random values The worst approach. Constant values will theoretically lead to all states representing the same unit. Random values are not good too. The vector state space is smaller than all possible values generating by random generator. Random value initialization is better than constant initialization, but majority of vectors have nonsense values.

Random values over training database Same approach as random values, but we overcomes nonsense values. Our space is limited to values existing in training database. Each mixture mean values are set to a random vector in database. All mixtures variances are set to global variance of the database. We initialize for example 30% of states to silence if database contains 30% of silence. It is disadvantage of this approach.

State splitting Iteratively trains EHMM. One selected state splits into two states. Means are set to be little bit different (adding and subtracting fraction of global mean variance) and EHMM is trained again. We tested two methods to select the candidate state to be split by the **largest amount of data** and by the **lowest log-likelihood**.

4 EHMM TRAINING AND RECOGNITION

We used the HTK tool [2] for EHMM training and recognition. This tool is developed for using HMM in speech processing. Baum-Welch forward-backward algorithm (*HRest*) was used for training (parameter estimation). Viterbi decoder was used for generating state aligned labelling (*HVite*). For more information about mathematical definitions and principles of HMM see for example [1].

5 RESULTS

We can not use standard evaluation methods because units are automatically determined (in comparison with phoneme recognition where we can compare generated and reference labels). For small EHMMs (to 5 emitting states) we can use some **visualization tool**. We can plot speech signal, spectra and generated state alignment. Coherency can be easily seen. Visualization starts to be difficult for larger EHMMs.

Another test is **listening** units belonging to one state. By this way, we can listen whether units of one state sound coherently. But this approach is quite time consuming, is subjective and can hardly detect different states covering similar type of units.

The most reliable way for evaluation of results are statistical methods. Labelled database is needed here. We calculate **mapping matrix** in which rows stand for phoneme string, columns represent states and cells contain numbers of phoneme strings corresponding to states. Having same number of states as phonemes, the matrix should be sparse (diagonal in ideal case). Calculation of the most frequent phoneme strings fall to state is another statistic.

Experiments were done on *Boston University Radio Corpus*[3] speaker *B2F*. Database was parameterized to 39 point vectors ($MFCC1-13$, $\Delta MFCC1-13$, $\Delta\Delta MFCC1-13$). Tested initialization methods and numbers of emitting states are shown in table 1. Mapping matrix, statistics and recognized streams of states were used for result comparison.

Type of initialization	Number of emitting state
Constant/random values	$2,3,4$
Random vectors from database	$2-80, 16/2, 32/4, 58/4, 100, 200$
State splitting by data	$2-80$
State splitting by log-likelihood	$2-60$

Table 1: Initialization methods and number of EHMM states. Notation y/x means initialization with x different return-probabilities in transition probability matrix.

5.1 EFFECT OF NUMBER OF STATES

The first experiment set was aimed to investigate into effect of number of states on states-to-phonemes assignment. Brief commentary for experimental results follows. EHMM with:

$2-5$ **states:** Independent on initialization method. State 1 labels silence, noise and unvoiced parts of speech. States $2,3,4$ label voiced parts of speech.

approximately 10 **states:** States divided into {vowels}, {silence}, {breathing, fricatives, plosives} and {nasal} groups.

approximately 20 **states:** States divided into {vowels}, {silence}, {breathing}, {fricatives}, {plosives} and {nasal} groups.

approximately 40 **states:** States divided into {vowels}, {silence}, {breathing}, {fricatives}, {plosives}, {glides}, {nasal} , {`s`, `z`} and {`r`} groups.

50 – *more* **states:** Similar as for 40 states. Groups cover approximately the same phonemes but in listening tests the groups sound more specific.

For number of states higher than 16, similar groups of phonemes appear. For example there were about 4 groups for vowels and 3 groups for {`s`,`z`} in 32 states model. Some groups have large intersection. Some of these facts are caused by coarticulation (the same phoneme in different contexts sounds differently). Listening test over states with similar groups (58 state EHMM) show, that each group is little bit different (longer, shorter, more/less voiced). Units in each group sound coherently.

5.2 RETURN TRANSITION PROBABILITY EFFECT

The second experiment set was aimed to investigate into the effect of return transition probabilities[4] initialization. Two sets with 32 and 58 states were tested. We tried to initialize EHMM with different in-state probabilities. We use high probability (it forces EHMM to stay in the same state longer time) and low probability (it forces EHMM to jump between more often).

Our results show that return-probability initialization has no effect on final recognition. Transition probabilities matrixes look similar in all cases. They are sparse for models with more states (it means that the EHMM is not fully connected anymore).

5.3 INITIALIZATION EFFECT

The third experiment set was aimed to model (prototype) initialization. Initialization approaches were described in section 3. Initialization has no effect for small state numbers (to 10). The comparison of approaches is shown in figure 2.

Differences between random vector and data driven selection are minimal. Log-likelihood driven approach is worse, because it splits states with lower likelihood. These states usually contain unvoiced speech, noise and plosives, so the result is more states for pauses than for vowels.

	vowels	glides	nasal	fricat.	stops	pause, breath
random vector	21	10	6	8	1	12
data driven	19	10	6	8	1	14
likelihood driven	15	10	4	7	3	19

Table 2: Number of states per groups of phonemes for different initialization approaches.

[4] Return-probability denotes probability $a_{22}, a_{33}, \ldots$

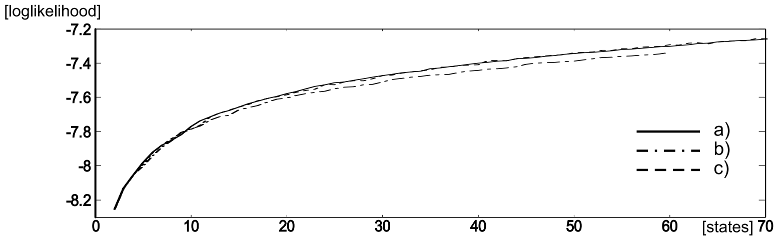

Figure 2: Log-likelihood dependency on number of states for a) data driven state splitting, b) log-likelihood driven state splitting, c) initialization by random vector from database. Y-axis is length normalized sum of log-likelihood over recognized database.

6 CONCLUSION

Ergodic hidden Markov model and some of its possible utilizations were presented. We aimed to investigate model initialization and training approaches such as state splitting.

This first experiments shows, that the EHMM is useful and can be used in automatic definition of speech units. We show, that transition probability matrix initialization has no important influence on model training. Initialization (training) method affects the quality of training. Data driven splitting method is the best one of presented training approaches.

In future, we will look for better formal result verification. Further experiments will be aimed to speaker and language independent testing and training. We will try to find a better method for model training. One of our goals is to train EHMM's with states as coherent with phonemes as possible. The speech group at FIT VUT is active in reliable phoneme recognition and we aim at using ALISP units as alternative approach. Another goal is to check different speech features such as TRAPS.

ACKNOWLEDGEMENTS

This research has been partially supported by Grant Agency of Czech Republic under project No. 102/02/0124 and by EC project Multi-modal meeting manager (M4), No. IST-2001-34485.

REFERENCES

[1] X. Huang, A. Acero, and H.W. Hon. *Spoken Language Processing*. Microsoft, Redmond, WA, USA, 1 edition, oct 2000.

[2] *HTK*. Homepage: http://htk.eng.cam.ac.uk/.

[3] M. Ostendorf, P.J. Price, and S. Shattuck Hufnagel. The Boston University radio news corpus. Technical report, Boston University, feb 1995.

[4] J. Černocký. *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification*. PhD thesis, Université Paris XI Orsay, dec 1998.