Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Telephone Conversation Speaker Diarization Using Mealy-HMMs

Itshak Lapidot[1], *Jean-François Bonastre*[2], *Samy Bengio*[3]

[1]Afeka Tel-Aviv College of Engineering, ACLP, Israel
[2]University of Avignon, LIA, France
[3]Google, Mountain View, USA

itshakl@afeka.ac.il, jean-francois.bonastre@univ-avignon.fr, bengio@google.com

Abstract

When *Hidden Markov Models* (HMMs) were first introduced, two competing representation models were proposed, the Moore model, with separate emission and transition distributions, which is commonly used in speech technologies, and the Mealy model, with a single emission-transition distribution. Since then the literature has mostly focused on the Moore model. In this paper, we would like to show the use of Mealy-HMMs for telephone conversation speaker diarization task. We present the Viterbi training and decoding for Mealy-HMMs and show that it yields similar performance compared to Moore-HMMs with a fewer number of parameters.

1. Introduction

Hidden Markov Models (HMMs) are well-known density estimators of sequences, often used in several domains like automatic speech recognition [1]. The most used variant of HMMs, also known as the Moore-HMM [2] is composed of two kinds of components: emission density functions $p(x_n|s_n)$, which model the density of observing x_n at time n while in state s_n, and transition probability distributions $p(s_n\,|s_{n-1}\,)$, which model the probability of going from a given state s_{n-1} at time $n-1$ to another state s_n at time n. This separation between emission and transition probabilities thus assumes that while in a given state s_n at time n, the probability of observing data x_n does not depend on the state the sequence was at the previous time step. Hence, if the distribution of x_n for sequences where the HMM went through state $s_{n-1} = i$ is very different from the distribution of x_n for sequences where the HMM went through state $s_{n-1} = j$, both will be blended into a single distribution at state s_n. This is due to the Markovian property of HMMs and the separation of emission and transition distributions, but can sometimes lead to bad representation.

In this paper, we would like to consider another representation of HMMs, introduced as Mealy-HMMs [3], where emission and transition distributions are merged into a single joint model. In the Mealy-HMM, the joint likelihood of being in state s_n at time n and emitting data x_n given the previous state was s_{n-1} is modeled by a single distribution of $p(x_n, s_n\,|s_{n-1}\,)$. Such distribution can then be used to model complete sequences similarly to Moore-HMMs, but without the downside of having two kinds of distributions to merge.

In order to overcome the difference in the nature of transition probabilities and emission densities several approaches have already been proposed in the literature. The so-called Fudge factor [4], for instance, is a poorly theoretically justified technical compensation which requires the estimation of a hyper-parameter on some development set, in order to compensate for the difference in variances. A better approach, dubbed the *Hidden-Distortion Model* (HDM) [5] was recently proposed. It also requires a hyper-parameter but it can be selected using the same training set as for all other parameters.

Speaker diarization is an unsupervised task where, for a given conversation, the goal is to determine "Who spoke when?". This is done by segmenting and clustering the conversation into homogeneous clusters, such that each cluster represents one speaker. If the number of speakers is not known it should be estimated from the conversation. Other events such as non-speech and overlapping speech should also be detected. Different tasks have their specific difficulties, such as meetings, shows, or telephone conversations. A variety of methods have been proposed to solve the speaker diarization task. A good overview can be found in [6].

In this paper we concentrate on a telephone conversation speaker diarization task. In this case the number of speakers is known *a-priori* and assumed to be always equal to 2. The difficulty comes from a relatively high speaker switching rate and the variety of different channels.

Telephone conversation speaker diarization can have important security applications as well as practical use in call centers where the the number of speakers is assumed to be always two. Nevertheless, there is not much work done on telephone conversation speaker diarization. The best reported results use *factor analysis* [7]. However, it is hard to compare their performance with the results in this paper as they are on a different database, the overlapping speech are not taken into account in [7] while we use a block to detect it, and the error is a part of the total error. When using iterative diarization systems, the Moore-HMM based speaker diarization system performance on the same database is not as good as HDM based systems [5], which showed better results. However the goal of this work is to show the potential of Mealy-HMMs and compare it with the same system using a Moore-HMM representation, as presented in [8]. The rest of the paper goes as follows: Section 2 formulates the training and decoding algorithms for Mealy-HMMs, using Viterbi; Section 3 describes the Mealy-HMM based diarization system; Section 4 presents the diarization result; and Section 5 concludes the paper.

2. Continuous Mealy-HMMs

While Mealy-HMMs were proposed several years ago, we haven't seen in the literature a precise Viterbi decoding algorithm or an EM parameter estimation algorithm for the Mealy-HMM, in particular for the continuous data case. The Baum-Welch algorithm was first presented in 2007 [9] for the discrete case.

Let us denote $p(x)$ the density of a continuous random vari-

able and $P(m)$ the probability of a discrete random variable.

Let us assume we have a system with K hidden states. We would like to define a joint transition-emission (TE) matrix

$$A(x_n) = [a_{qk}(x_n)]_{q,k\in\{1,\ldots,K\}}$$

where

$$a_{qk}(x_n) = p(x_n, s_n = q | s_{n-1} = k)$$

is the probability of being at time n in state q and observation $x_n \in \mathbb{R}^{d\times 1}$ given the previous state was k. Let $X_{n_1}^{n_2} = [x(n_1), x(n_1+1), \ldots, x(n_2)]$ be a partial sequence ($X = X_1^N$ is the full sequence). The assumption is that the TE density depends only on the previous state, i.e.,

$$\begin{aligned} p\big(x_n, s_n = q \,\big|\, X_1^{n-1}, X_{n+1}^N, s_{n-1} = k, \\ \{s_m = k_m\}_{m\notin\{n-1,n\}}\big) \\ = p(x_n, s_n = q \,|\, s_{n-1} = k) \end{aligned} .$$

The constraint is that the integration over the sum of each column of $A(x_n)$ gives 1,

$$\forall k \bullet \int \sum\nolimits_{q=1}^{K} a_{qk}(x_n) dx_n = 1.$$

In such formulation we do not assume that the emission and transition probabilities are statistically independent, given the model.

During all the discussion we will assume that

$$\forall q,k \bullet a_{qk}(x_n) = \sum\nolimits_{m=1}^{M_{qk}} \omega_{qk}^m \mathcal{N}\left(x_n; \mu_{qk}^m, \Sigma_{qk}^m\right)$$

is a linear combination of Gaussian mixture components, such that

$$\sum\nolimits_{q=1}^{K} \sum\nolimits_{m=1}^{M_{qk}} \omega_{qk}^m = 1,$$

sum over **all** the mixture component weights for all TE functions which transit from state k equals 1. We denote ω as the mixture weight; $\mu \in \mathbb{R}^{d\times 1}$ as the mixture mean vector; $\Sigma \in \mathbb{R}^{d\times d}$ as the mixture covariance matrix; and M_{qk} as the number of mixture components.

In addition we define prior TE density functions to be in state k with the first observation,

$$\Pi(x_1) = [\pi_1(x_1), \ldots, \pi_K(x_1)]^T,$$

$$\pi_k(x_1) = p(x_1 | s_1 = k)$$

such that

$$\int \sum\nolimits_{k=1}^{K} \pi_k(x_1) dx_1 = 1.$$

From now on, we will assume for simplicity that $\forall q,k \bullet M_{qk} = M_k = M$.

We now enumerate two problems to be solved for continuous Mealy-HMMs for the task of speaker diarization:

1. Given the model $\mathcal{M} = \{A, \Pi\}$, find the path S^* which maximizes the log-likelihood of a sequence of data samples $X = \{x_1, \ldots, x_N\}$ and the sequence of states $S = \{s_1, \ldots, s_N\}$:

$$\begin{aligned} L(X, S^* | \mathcal{M}) &= \log p(X, S^* | \mathcal{M}) \\ &= \max_{\{s_n\}} \left\{ \tilde{\pi}_{s_1}(x_1) + \sum_{n=2}^{N} \tilde{a}_{s_n s_{n-1}}(x_n) \right\} \end{aligned} \quad (1)$$

where $\tilde{\pi}_k(x_1) = \log \pi_k(x_1)$ and $\tilde{a}_{qk}(x) = \log a_{qk}(x)$.
This problem can be solved using the well known Viterbi algorithm.

2. The parameter estimation problem (we will formulate it here only using Viterbi as well): given the data $\mathbf{X} = \{X^i\}_{i=1}^I$, consisting of I sequences each one of length N_i, and the model parameters $\mathcal{M}$, we would like to find a new model $\hat{\mathcal{M}}$ which maximizes the log-likelihood $L(X, S^* | \mathcal{M})$.

2.1. Viterbi Algorithm

To find the best path we have to define two variables: $\delta(k,n) = \max_{S_1^{n-1}} p\left(S_1^{n-1}, s_n = k, X_1^n \middle| \mathcal{M}\right)$ where $S_{n_1}^{n_2} = \{s_{n_1}, s_{n_1+1}, \ldots, s_{n_2}\}$, $\delta(n) = [\delta(1,n), \delta(2,n), \ldots, \delta(K,n)]^T$, and $\psi(k,n) = {S^*}_1^{n-1}(n-1)$ where ${S^*}_1^{n-1} = \arg\max_{S_1^{n-1}} p\left(S_1^{n-1}, s_n = k, X_1^n \middle| \mathcal{M}\right)$, and $\psi(n) = [\psi(1,n), \psi(2,n), \ldots, \psi(K,n)]^T$.

Initialization:

$$\begin{aligned} \delta(1) &= \Pi(x_1) \\ \psi(1) &= [0, 0, \ldots, 0]^T \end{aligned} \quad (2)$$

Recursion:

$$\begin{aligned} \delta(k,n) &= \max\left\{A_{k\cdot}(x_n)^\circ \delta^T(n-1)\right\} \\ \psi(k,n) &= \arg\max\left\{A_{k\cdot}(x_n)^\circ \delta^T(n-1)\right\} \end{aligned} \quad 2 \le n \le N \quad (3)$$

where $A_{k\cdot}(x)$ is the k^{th} row of $A(x)$ and the operator $\circ$ is the Hadamard product.

Termination:

$$\begin{aligned} p(X, S^* | \mathcal{M}) &= \max \delta(N) \\ s_N^* &= \arg\max \delta(N) \end{aligned} \quad (4)$$

Backtracking:

$$s_{n-1}^* = \psi(s_n^*, n) \quad n = N-1, N-2, \ldots, 1 \quad (5)$$

2.2. Parameters Estimation

In this section we present the estimation procedure of the continuous Mealy-HMM parameters, including the TE initial vector and the TE matrix.

2.2.1. Viterbi Parameter Estimation

In order to apply the Viterbi algorithm, in addition to the data $\mathbf{X} = \{X^i\}_{i=1}^I$ and initial model $\mathcal{M}$, the sequences of states $\mathbf{S} = \{S^i\}_{i=1}^I = \{s_n^i = k_n^i\}_{\substack{i=1,\ldots,I \\ n=1,\ldots N_i}}$, provided by the Viterbi decoding algorithm are also given. The log-likelihood of the data and the state sequences given the model is:

$$
\begin{aligned}
&L\left(\mathbf{X},\mathbf{S}|\mathcal{M}\right)=\log\left(p\left(\mathbf{X},\mathbf{S}|\mathcal{M}\right)\right)\\
&=\log\left(\prod_{i=1}^{I}p\left(x_1^i\middle|\mathcal{Q}_{s_1^i}\right)\cdot\prod_{n=2}^{N_i}p\left(x_n\middle|\mathcal{M}_{s_n^i s_{n-1}^i}\right)\right)\\
&=\sum_{i=1}^{I}\left[\log p\left(x_1^i\middle|\mathcal{Q}_{s_1^i}\right)+\sum_{n=2}^{N_i}\log p\left(x_n\middle|\mathcal{M}_{s_n^i s_{n-1}^i}\right)\right] \qquad (6)\\
&=\sum_{k=1}^{K}\sum_{\left\{n:s_n^i=k\right\}}\log p\left(x_n^i\middle|\mathcal{Q}_k\right)\\
&+\sum_{q,k\in\{1,\ldots,K\}}\ \sum_{\left\{n:\left(s_n^i,s_{n-1}^i\right)=(q,k)\right\}}\log p\left(x_n|\mathcal{M}_{qk}\right)
\end{aligned}
$$

where $\mathcal{M}_{qk}$ and $\mathcal{Q}_k$ are the models of the TE density function $a_{qk}(x)$ and TE initial density function $\pi_k(x)$ respectively.

Let

$$X_{qk}=\left\{x_n^i:\left(s_n^i,s_{n-1}^i\right)=(q,k)\right\},$$

be all the data associated with the q^{th} TE function from the k^{th} state, and n_{qk} be the cardinality of the X_{qk} set. We will also define $X_k=\bigcup_{q=1}^{K}X_{qk}$ as the data associated with the k^{th} state, and $n_k=\sum_{q=1}^{K}n_{qk}$ as a cardinality of X_k (we define the transition-emission prior $P_{qk}=\frac{n_{qk}}{n_k}$). For each state the model is the union of the models of all TE functions, $\mathcal{M}_k=\{\mathcal{M}_{1k},\ldots,\mathcal{M}_{Kk}\}$. The model of each state can be trained separately with the data associated by the Viterbi decoding $\{X_k\}_{k=1}^{K}$. It is important to clarify that if the GMM model trained by the data X_{qk} is

$$\mathcal{G}_{qk}=\left\{\left\{\omega_{qk}^m\right\},\left\{\mu_{qk}^m\right\},\left\{\Sigma_{qk}^m\right\}\right\}$$

then

$$\mathcal{M}_{qk}=\left\{\left\{P_{qk}\cdot\omega_{qk}^m\right\},\left\{\mu_{qk}^m\right\},\left\{\Sigma_{qk}^m\right\}\right\}$$

and it is not a model of a *pdf* as it integrates to P_{qk}. On the other hand $\mathcal{M}_k$ is a model of a *pdf*.

2.3. Relation Between Mealy-HMMs and Moore-HMMs

In the past, equivalences between Mealy and Moore types of HMMs were considered, but it was mostly done for the discrete case. We would like to consider two important issues: equivalent models and minimal models.

Let $\mathbb{X}^*$ be the set of all possible sequences, e.g., $\mathbb{X}=\{0,1\}$ then $\mathbb{X}^*=\{\emptyset,0,1,00,01,\ldots\}$, and string probabilities $\mathcal{P}:\mathbb{X}^*\mapsto[0,1]$ then:

Definition of Equivalent Models: Two HMMs (both Mealy, both Moore, or one Mealy and another Moore) with string probabilities $\mathcal{P}$ and $\mathcal{P}'$ respectively, are equivalent, if $\mathcal{P}=\mathcal{P}'$.

Mealy Minimal Model Definition: A Mealy-HMM with rank K is called minimal if for any other equivalent Mealy-HMM with rank K', $K\leq K'$.

Moore Minimal Model Definition: A Moore-HMM with rank K is called minimal if for any other equivalent Moore-HMM with rank K', $K\leq K'$.

While finding a minimal model is still an open question in general, for discrete HMMs it was shown that the "expressive power" of Mealy- and Moore-HMMs are the same, meaning that for each Moore-HMM there exists a Mealy-HMM equivalent model and vice versa: for every Mealy-HMM there exists a Moore-HMM equivalent model. It has also been shown that the order of a minimal Mealy model does not exceed the order of a minimal Moore model [2]. It can be shown that the "expressive power" of the continuous Mealy- and Moore-HMMs are also the same and that the order of a minimal Mealy model does not exceed the order of a minimal Moore model. It can be shown that the number of states in the Moore-HMM can be up to K^2 in order to be equivalent to a K-state Mealy-HMM. The state have to be grouped into k-state groups to represent one state in the Mealy-HMM, as shown in Figure 1.

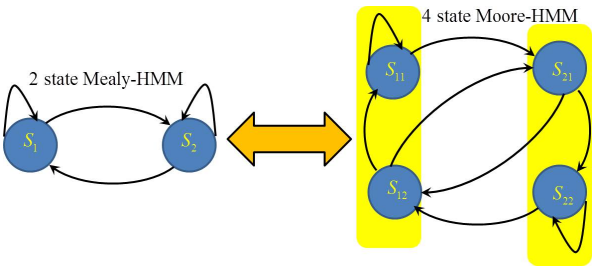

Figure 1: *Two state Mealy-HMM represented as 4-state Moore-HMM.*

2.3.1. Moore-HMM vs Mealy-HMM Discussion

This section discusses the relation between Moore- and Mealy-HMMs with respect to the number of parameters. We consider two cases:

1st : The true model is a K-state Moore-HMM but we model it with a Mealy-HMM. A K-state Moore-HMM can be replaced by a K-state Mealy-HMM with the same order GMM on the transitions, i.e., replacing one M-component GMM in a state by K GMMs with M components each in the transitions. This means that we need to estimate $K-1$ times more parameters, however the Mealy-HMM is equivalent to the Moore-HMM. $K\times(K-1)$ transition probabilities have to be estimated in the Moore-HMM while in Mealy-HMM they are part of the GMM parameters, but still required.

2nd : The true model is a K-state Mealy-HMM but we model it with a Moore-HMM. In this case, the number of parameters is reduced, but the model is not equivalent to the original Mealy-HMM. This will most likely damage the performance. In order to correct for this, we need to increase the number of states, but we do not know how many states should be added, up to K^2 states. In this case, the number of GMM parameters becomes the same, but the number of transitions becomes K^{2^2}. Most of the transitions should tend to zero, and only K^2 should have meaningful values. After training of the K^2-state Moore-HMM, if it is trained for segmentation purpose, the states should be grouped into K groups of K states each. Each group represents one event (one state in the Mealy-HMM). We see that a Moore-HMM does not increase dramatically the number of estimated parameters (in the transition matrix only) if any, but adds a lot of uncertainty, as we do not know how many states are required and which states are related to the same event.

2.3.2. Moore-HMM vs Mealy-HMM for Diarization

In the case of speaker diarization, we can imagine that the start of speech (the transition) may differ by speaking after a silence

or just after another speaker. In the first case, the speech may be more calm and quite, while after another speaker it might be louder. These differences in the speech should appear in the model representation. In our model we have one hyper-state for each speaker (K speakers) plus one hyper-state for the non-speech, Section 3.1. Three options are possible. Two options with Moore-HMM and one with Mealy-HMM:

1st : The number of states in the Moore-HMM is equal to the number of speakers plus one $(K+1)$. In order to describe all the possible transition effects, a GMM with many mixture components is required.

2nd : The transition from one speaker (or non-speech) to another will first pass through extra state which will account for the transition effects. This way the speaker hyper-state GMM can have less mixture components, however, additional $(K+1) \times K$ states for transitions should be added. The total number of states becomes $(K+1)^2$ and it makes the transition matrix to be $(K+1)^2 \times (K+1)^2$. Although, the transition matrix is sparse, but still the segmentation is much more time consuming.

3rd : Using Mealy-HMM with $K+1$ hyper-states. As the GMMs are on the transitions, and not in the states, the transition effects are modeled in a natural way. No extra state is required; the speaker model may have less mixture components; the segmentation is even simpler than in the $(K+1)$-state Moore-HMM.

We see that a Mealy-HMM has only advantages over the Moore-HMM. In Section 4 we will show that Mealy-HMM achieves a bit better results than Moore-HMM with fewer total number of mixture components. If the number of speakers is not known in advance, and transition states are provided in the Moore-HMM, a merging process of two speakers becomes less trivial as it is not done by replacing two states by one state with merged data, but also all the transition states to the states and from the states should be replaced and retrained and it is much more complicated.

3. Diarization System

In this section we describe the telephone conversation speaker diarization system for both Moore- and Mealy-HMM cases.

3.1. Speaker Diarization System

The baseline diarization system corresponds to [5] where more details could be found. Figure 2 shows the block diagram of the system. 12 *mel-frequency cepstral coefficient* (MFCC) features from 20 ms windows are extracted each 10 ms. A simple threshold *voice activity detection* (VAD) is applied for initial speech/non-speech segmentation. In parallel, an overlapping speech detection is performed. The overlapping speech detector is based on maximum *a-posteriori* estimator of the wave form entropy, which is estimated each 100 ms [10]. The detected overlapped speech is taken out of the conversation, before performing speaker model initialization using *weighted-segmental K-means* (WSKM) [8]. The diarization process itself is based on an 3-hyper-state fixed-duration Mealy- or Moore-HMM (Sections 3.2 and 3.3), for two speakers and non-speech classes. Each hyper-state model has tied states of 200 ms for the first 5 iterations and only 100 ms tied states for the last iteration. The state models we used for Moore-HMM are 21 Gaussian component GMMs with full covariance matrices, as this gave the best diarization performances [8], and 24 Gaussian component GMMs in order to have an overall same number of Gaussians as in the best Mealy-HMM system. In the Mealy-HMM, the number of mixtures in the transition states and in the hyper-states were varied in order to obtain the best configuration. The transition matrix is initialized using the initial segmentation provided from the VAD and the WSKM. The Viterbi decoding gives a new segmentation and clustering, which is used for the subsequent retraining iteration.

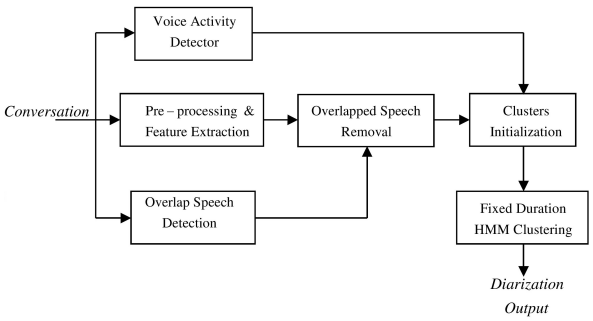

Figure 2: *Baseline diarization system.*

In order to compare the Moore-HMM based speaker diarization system with a Mealy-HMM based speaker diarization system we keep the same configuration but we replace the Fixed-duration Moore-HMM block (Section 3.2 and Figure 3) by a Fixed-duration Mealy-HMM block, as explained in Section 3.3.

3.2. Fixed-Duration Moore-HMM

Figure 3 shows an example of a 3-hyper-state fix-duration Moore-HMM. When the system enters into one hyper-state, it has to stay in this hyper-state for a predefined number of frames, τ (20 in our case). At the last frame of the fixed-duration segment, the system can transit to the first state of any hyper-state (including returning to the first state of the same hyper-state). At each iteration, the best path is found using Viterbi decoding.

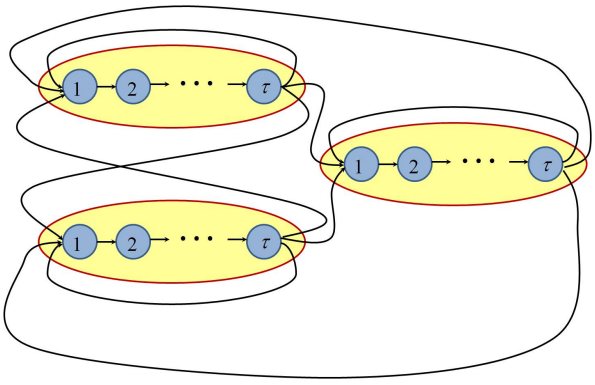

Figure 3: *3-state fixed duration Moore-HMM.*

For K hyper-state Moore-HMM, the transition matrix is

composed of $K \times K$ blocks:

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1K} \\ A_{21} & A_{22} & \cdots & A_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ A_{K1} & A_{K2} & \cdots & A_{KK} \end{pmatrix} \tag{7}$$

Each block A_{qk} is a $\tau \times \tau$ matrix (τ is the number of states in each hyper-state). Each matrix on the main diagonal is the internal hyper-state transition matrix and it contains only zeros except ones on the diagonal below the main diagonal and $P(k|k)$, the probability of returning to the first state of the hyper-state (self-loop on the hyper-state level), at the right element of the first row (eq. 8). Each matrix out of the main diagonal contains only zeros except the right element of the first row, which is the transition probability to go to hyper-state q from hyper-state k, $P(q|k)$ (eq. 9).

$$A_{kk} = \begin{pmatrix} 0 & \cdots & 0 & P(k|k) \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix} \in \mathbb{R}^{\tau \times \tau} \tag{8}$$

$$A_{qk} = \begin{pmatrix} 0 & \cdots & 0 & P(q|k) \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 \end{pmatrix} \in \mathbb{R}^{\tau \times \tau} \tag{9}$$

The probabilities at position $(1, \tau)$ are trained using the Viterbi statistics.

3.3. Fixed-Duration Mealy-HMM

Figure 4 shows an example of a 3-hyper-state fixed-duration Mealy-HMM. Unlike with Moore-HMMs, a Mealy-HMM system enters into a hyper-state through one TE state (pink circle) from each hyper-state, and then has to stay in this hyper-state for a predefined number of frames, $\tau - 1$ (19 in our case). At the last frame of the fixed-duration segment, the system transits to the first state of any hyper-state via the TE state (including returning to the first state of the same hyper-state). Inside the hyper-state, all the TE density functions are tied, i.e., share all their parameters. At each iteration, the best path is found using Viterbi decoding. If it is not the final iteration, the models are retrained with the Viterbi training algorithm (Section 2.2.1).

4. Experiments and Results

In this section we describe the experimental setup and present the results of the Mealy-HMM based system as compared to a Moore-HMM-based system.

4.1. Diarization Error Rate (DER)

The results are presented in terms of *diarization error rates* (DER) [11] with a non-scoring collar of $0.5sec$ around the changing points, i.e., $0.25sec$ on each side of the changing points. The DER is defined as follows:

$$DER = 100 \frac{\sum_{s=1}^{S} dur(s) \cdot \max(N_{Ref}(s), N_{Sys}(s)) - N_{Cor}(s))}{\sum_{s=1}^{S} dur(s) \cdot N_{Ref}(s)} \tag{10}$$

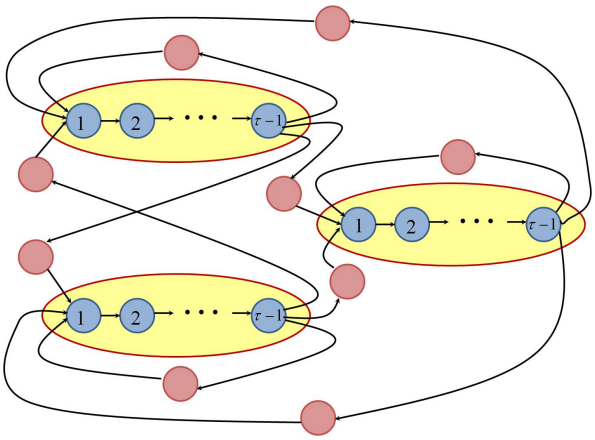

Figure 4: *3-state fixed duration Mealy-HMM.*

Given S speech segments in the conversation, the DER is calculated according to the following notation:

- $dur\,(s)$ - Duration of segment s.
- $N_{Ref}\,(s)$ - The number of speakers assigned to segment s.
- $N_{Sys}\,(s)$ - The number of speakers assigned by the system to segment s.
- $N_{Cor}\,(s)$ - The number of speakers assigned by the system to segment s which actually takes part in s.

Since the indexing of the speakers by the system does not use any prior knowledge about their identity, the DER should be calculated for all possible permutations of the speaker indices, and the minimum obtained DER is taken.

4.2. Database

We used a subsets of 108 conversations extracted from LDC America Call Home English language corpus [12]. The database is sampled at 8 kHz in a 2 channel μ-law format. The channels were summed to generate a two-speaker conversation. Only the transcribed part of each conversation was taken, resulting in about 10 minutes per conversation.

4.3. Experiments

We first examine the Moore-HMM-based diarization system with 21 and 24 Gaussian components per state. These systems are the baseline to be compared to the Mealy-HMM based diarization system and the DER results are presented in Table 1. We can see that both configurations give the same DER, however the second configuration has 9 Gaussian components more. With 27 Gaussians per state the results are significantly worse.

Table 1: *Baseline Moore-HMM diarization results.*

Gaussians per hyper-state	**21**	**24**	**27**
DER [%]	21.73	21.45	24.43

Different configurations of Mealy-HMM were then examined, mostly varying the number of Gaussian components in the

hyper-state and the TE distribution functions. The results are summarized in Table 2. The first row specifies the number of Gaussian components in the hyper-state distribution functions, while the first column specifies the number of Gaussian component in the TE distribution functions. We can see that the results improve as the number of Gaussian components in the hyper-state increases and the best results are achieved with 21 and 24 components. For the TE, if the number of mixture components in the hyper-state is sufficiently large, 1 Gaussian can give DER close to the best DER, achieved for 3 Gaussian components only. This is probably due to the fact that only once in 20 time stamps one of the 3 TE is entered, which means that TE distribution functions are trained using only 5% of the data. The configuration of 16 Gaussian components in the hyper-states and one Gaussian in the TE achieves a DER similar to the baseline system with 6 components less. Similar DER and same number of Gaussian components are in the configuration of 10 Gaussian components in the hyper-state with 3 Gaussian components in the TE. 6 Gaussian components, both in hyper-state and TE has a complexity of Moore-HMM with 24 Gaussian components and similar DER. The configuration of 21 Gaussian components in the hyper-state and 1 Gaussian components in the TE has the same number of Gaussian components as the baseline system with 24 Gaussian components and achieves a slight reduction in the DER (about 2.5% relative improvement). When we use 24 Gaussian components in the hyper-state and 3 Gaussian components in the TE, relative improvement is about 8.9%. The complexity of such Mealy-HMM is equivalent to Moore-HMM with 33 Gaussian components. However, in Moore-HMM, when the number of Gaussian components per state goes above 24 the DER becomes higher (as can be seen in the case of 27 Gaussian components, in Table 1).

Table 2: *Mealy-HMM diarization results in terms of DER [%], for a given number of Gaussians in the hyper-state (rows) and in the TE (columns).*

	2	**3**	**6**	**10**	**16**	**21**	**24**
1	28.67	27.11	24.33	22.20	21.23	20.91	20.73
2	28.82	27.02	24.57	22.21	21.02	20.42	20.22
3	-	25.94	23.59	21.61	20.41	19.69	19.54
6	-	-	21.61	24.43	23.26	21.26	21.91
10	-	-	-	-	-	32.94	-

5. Conclusion

In this work we presented training and decoding algorithms for Mealy-HMMs as well as an application to speaker diarization. The advantage of the Mealy-HMM structure given the fixed-duration constraint is limited due to the fact that Mealy-HMMs have an impact only on the transitions from one hyper-state to another. Nevertheless, we showed that we achieve the same performance as Moore-HMMs with a smaller model and better performance with a Mealy-HMM which has the same number of Gaussian components. When we use a Mealy-HMM with higher number of Gaussian components than the best Moore-HMM configuration, the relative improvement is about 8.9%.

The approach is not limited to the case of two speakers and it will be interesting to test Mealy-HMMs on conversations with more speakers. Estimation of the number of speakers is not within the scope of the paper, but it would be interesting to test it on the presented model. Although Mealy-HMMs should naturally balance the emissions and the transitions, as they are blended into one distribution, in the Viterbi training case the decoupling is possible. So a full optimization could be done in order to obtain a better clustering, for example with the "fudge factor" (as it is done in HDMs).

6. References

[1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech processing," *Proceeding of the IEEE*, vol. 77, no. 2, February 1989.

[2] B. Vanluyten, J.C. Willems, and B. De Moor, "Equivalence of state representations for hidden markov models," *Systems and Control Letters*, vol. 57, no. 5, pp. 410 – 419, 2008.

[3] G. Mealy, "A method for synthesizing sequential circuits," *Bell System Technical Journal*, pp. 1054–1079, 1955.

[4] B. Lecouteux, G. Linares, Y. Esteve, and G. Gravier, "Generalized driven decoding for speech recognition system combination," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-april 4 2008, pp. 1549 –1552.

[5] I Lapidot and J.-F. Bonastre, "Generalized viterbi-based models for time-series segmentation applied to speaker diarization," in *ODYSSEY 2012 -The Speaker and Language Recognition Workshop*, 2012.

[6] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356 –370, feb. 2012.

[7] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059 –1070, dec. 2010.

[8] O. Ben-Harush, O. Ben-Harush, I. Lapidot, and H. Guterman, "Initialization of iterative-based speaker diarization systems for telephone conversations," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 414 –425, feb. 2012.

[9] B. Vanluyten, J.C. Willems, and B. De Moor, "A new approach for the identification of hidden markov models," in *Decision and Control, 2007 46th IEEE Conference on*, dec. 2007, pp. 4901 –4905.

[10] O Ben-Harush, I Lapidot, and H Guterman, "Entropy based overlapped speech detection as a pre-processing stage for speaker diarization," in *Proceedings of Interspeech 2009*, 2009.

[11] "Diarization error rate," Available: http://www.xavieranguera.com/phdthesis/node108.html.

[12] "Liguistic data consortium," LDC97S42, Catalog, 1997, Available: http://www.ldc.upenn.edu/Catalog.