Data-Derived Models for Segmentation with Application to Surgical Assessment and Training

Balakrishnan Varadarajan[1], Carol Reiley[2], Henry Lin[2], Sanjeev Khudanpur[1,2], and Gregory Hager[2]

[1] Department of Electrical and Computer Engineering
[2] Department of Computer Science, Johns Hopkins University,
Baltimore, MD 21218, USA
{bvarada2,creiley,hcl,khudanpur,hager}@jhu.edu

Abstract. This paper addresses automatic skill assessment in robotic minimally invasive surgery. Hidden Markov models (HMMs) are developed for individual surgical gestures (or *surgemes*) that comprise a typical bench-top surgical training task. It is known that such HMMs can be used to recognize and segment surgemes in previously unseen trials [1]. Here, the topology of each surgeme HMM is designed in a data-driven manner, mixing trials from multiple surgeons with varying skill levels, resulting in HMM states that model skill-specific *sub-gestures*. The sequence of HMM states visited while performing a surgeme are therefore indicative of the surgeon's skill level. This expectation is confirmed by the average edit distance between the state-level "transcripts" of the same surgeme performed by two surgeons with different expertise levels. Some surgemes are further shown to be more indicative of skill than others.

1 Automatic Skill Assessment in Robotic Surgery

Robotic minimally invasive surgery (RMIS) has experienced rapid development and growth over the past decade, and the da Vinci robotic surgery system has emerged as the leader in RMIS [2]. Training for RMIS has often been cited as challenging, even for experienced surgeons [3]. One approach to overcome this challenge is develop techniques for automatic assessment of surgical skills during the performance of benchmark tasks, such as suturing or knot-tying, that simulate live tasks used for clinical skill evaluation [4]. This paper presents such techniques based on gesture recognition using hidden Markov models (HMMs).

RMIS is uniquely amenable to automatic skill assessment. The robot functions as a *measurement tool* for dexterous motion. As part of its run-time system, the da Vinci exposes an application programming interface (API) which provides accurate and detailed kinematic motion measurements, including the surgeon console "master" manipulators and all patient-side tools. We use these measurements to recognize individual surgical gestures [1]. Using both surgeon- and patient-side kinematics may seem redundant. But since one may carry some information that the other doesn't, (e.g intended v/s actual tool motion), we use both, and apply data-driven dimensionality reduction techniques to remove such redundancies.

G.-Z. Yang et al. (Eds.): MICCAI 2009, Part I, LNCS 5761, pp. 426–434, 2009.

Dosis et al [5] have used hidden Markov models to model hand manipulations and to classify simple surgical tasks. Richards et al [6] have demonstrated that force/torque signatures may be used in RMIS for two-way skill classification. Rosen et al [7] have used HMMs to model tool-tissue interactions in laparoscopic surgery; a seperate HMM for each skill level was trained using a pool of surgeons, and a *statistical distance* between these HMMs was shown to correlate well with the learning curve of these trainee surgeons. In these and other reported efforts, the automatic assessment is for entire trials, while the work presented here assesses *finer grained* segments, namely individual surgical gestures.

Lin et al [8] have used linear discriminant analysis (LDA) to project the high-dimensional kinematic measurements from the da Vinci API to three or four dimensions, and used a Bayes' classifier to segment surgical gestures from the low-dimensional signal. Reiley et al [1] replace their Bayes classifier with a 3-state left-to-right HMM for each gesture, and demonstrate improved accuracy on unseen users. The work presented here improves upon [1] by performing LDA to discriminate between the kinematical signal of *sub-gestures* – modeled by individual HMM states – rather than between the signal of entire gestures.

The distinguishing contribution of this work is the application of the HMM methodology to gesture-specific skill assessment. A data-driven algorithm is used to design the HMM topology for each gesture. As a consequence, in addition to automatic detection and segmentation of surgical gestures, one is able to compare individual gestures of expert, intermediate and novice surgeons in a quantitative manner. For instance, some gestures in a suturing task, such as navigating a needle through the tissue, are demonstrated to be more indicative of expertise than others, such as pulling the thread. Such fine grained assessment can ultimately lead to better automatic surgical assessment and training methods.

This paper is organized as follows. We begin in Section 2 with a background review of the suturing task and the use of HMMs for gesture recognition and segmentation. We then describe the two technical novelties in the use of HMMs, namely state-specific LDA and data-derived HMM topologies, in Section 3. This leads to improved gesture recognition accuracies. In Section 4, we demonstrate how paths through the HMM state space are indicative of the expertise with which the gesture has been performed, leading to the main contribution of the paper: a framework for automatic, gesture-level surgical skill assessment.

2 Surgical Gesture Recognition Using HMMs

2.1 The Surgeme Recognition Experimental Setup

Kinematic Data Recordings: We recorded the kinematic measurements from 2 expert, 3 intermediate and 3 novice surgeons performing a bench-top suturing task—four stitches along a line—on the teleoperated da Vinci surgical system. The average duration of a trial is 2 minutes, and the video and kinematic data are recorded at 30 frames per second. The kinematic measurements include position, velocity, etc. from both the surgeon- and patient-side manipulators for a total of 78 motion variables. We use $\{y_t, t = 1, 2, \ldots, T\}$ to denote the sequence of kinematic measurements for a trial, with $y_t \in \mathbb{R}^{78}$ and $T \approx 3400$. A total of 30 trials were recorded, roughly four from each of the eight surgeons.

Manual Labeling of Surgemes: Each trial was manually segmented into semantically "atomic" gestures, based on the eleven-symbol vocabulary proposed by [1]. Following their terminology, we will call each gesture a *surgeme*. Typical surgemes include, for instance, (i) positioning the needle for insertion with the right hand, (ii) inserting the needle through the tissue till it comes out where desired, (iii) reaching for the needle-tip with the left hand, (iv) pulling the suture with the left hand, etc. We use $\{\sigma_{[i]}, i = 1, 2, \ldots, k\}$ to denote the surgeme label-sequence of a trial, with $\sigma_{[i]} \in \{1, \ldots, 11\}$ and $k \approx 20$, and $[b_i, e_i]$ the begin- and end-time of $\sigma_{[i]}$, $1 \leq b_i < e_i \leq T$. Note that $b_1 = 1$, $b_{i+1} = e_i + 1$, $e_k = T$.

The Surgeme Recognition Task: Given a partition of the 30 trials into *training* and *test* trials, the surgeme recognition task is to automatically assign to each trial in the test partition a surgeme transcript $\{\hat{\sigma}_{[i]}, i = 1, 2, \ldots, \hat{k}\}$ and time-marks $[\hat{b}_i, \hat{e}_i]$. Trials in the training partition are used to train the HMMs, as described below. We report results with three different training/test partitions.

Setup I: Of the 30 trials, 8 have some minor errors by the surgeons during suturing. These are excluded altogether in Setup I. Leave-one-out cross-validation is carried out with the remaining 22 trials, so that each trial is once in the test partition. The test results of all 22 folds (22 trials) are aggregated.

Setup II: The training partition in Setup II comprises the 22 "good" trials, while the test partition comprises only the 8 "imperfect" trials.

Setup III: User-disjoint partitions of the 30 trials are created in Setup III. An eight-fold cross validation akin to Setup I is carried out, except that in each fold, all the trials of 1 surgeon are in the test partition and all trials of the remaining 7 surgeons are in training. Test results of all 30 trials are aggregated.

Setup I is relatively the easiest, with 22 good test trials and the surgeon of each test trial seen in training. Setup II is harder, with seen surgeons but with test trials that have some visible errors, a situation not dissimilar from recognition of slightly disfluent speech. Setup II is most similar to the multiple-user results in [1, Table 3], with which we make direct comparisons. Setup III is the hardest, because all trials of the test surgeon have also been removed from training.

Recognition accuracy is measured as the fraction of kinematic frames that are assigned the correct surgeme label by an automatic system. Formally,

$$\text{Accuracy of test trial } \{y_1, \ldots, y_T\} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\left(\sigma_t = \hat{\sigma}_t\right), \tag{1}$$

where $\sigma_t = \sigma_{[i]}$ for all $t \in [b_i, e_i]$ and $\hat{\sigma}_t = \hat{\sigma}_{[i]}$ for all $t \in [\hat{b}_i, \hat{e}_i]$. This measures the goodness of both the labels *and* the segmentation proposed by $\{\hat{\sigma}_t\}$.

2.2 HMM-Based Surgeme Recognition

Dimensionality Reduction: Before surgeme recognition, the 78-dimensional kinematic data are reduced to $d \ll 78$ dimensions via LDA [9]. Specifically, each block of $2p+1$ frames in the training partition is converted into a data-label pair $\big([y_{t-p}^{\mathrm{T}} \cdots y_{t-1}^{\mathrm{T}}\, y_t^{\mathrm{T}}\, y_{t+1}^{\mathrm{T}} \cdots y_{t+p}^{\mathrm{T}}]^{\mathrm{T}},\, \sigma_t\big)$, and a $d \times 78(2p+1)$ projection matrix $\mathbf{A}$ is computed that maximizes the ratio of between- and within-surgeme scatter of the *projected* data $x_t = \mathbf{A}[y_{t-p}^{\mathrm{T}} \cdots y_t^{\mathrm{T}} \cdots y_{t+p}^{\mathrm{T}}]^{\mathrm{T}}$. Typically, $p = 5$ and d is 3 to 10. The $\{x_t\}$ are used everywhere subsequently, instead of $\{y_t\}$.

Surgeme Modeling: The likelihood of the kinematic signal $\{x_t, t = b_i, \ldots, e_i\}$ of a surgeme $\sigma_{[i]} = \sigma$ is modeled via a HMM as

$$P_\sigma(x_{b_i}, \ldots, x_{e_i}) = \sum_{s_{b_i} \in \mathcal{S}_\sigma} \sum_{s_{b_i+1} \in \mathcal{S}_\sigma} \cdots \sum_{s_{e_i} \in \mathcal{S}_\sigma} \prod_{t=b_i}^{e_i} p(s_t | s_{t-1}) \mathcal{N}(x_t\,;\, \mu_{s_t}, \Sigma_{s_t}), \quad (2)$$

where $\mathcal{S}_\sigma$ denotes the *hidden states* of the model for surgeme σ, $p(s|s')$ are the transition probabilities between these states, and $\mathcal{N}(\cdot\,;\, \mu_s, \Sigma_s)$ is a multivariate Gaussian density with mean μ_s and covariance Σ_s associated with state $s \in \mathcal{S}_\sigma$.

Parameter Estimation: Kinematic data from all training samples of a surgeme σ are modeled by the same HMM (with states $\mathcal{S}_\sigma$), and each surgeme is modeled by a different HMM. Model parameters are chosen to maximize the likelihood (2) of the training data $\{x_t\}$ via the standard Baum-Welch algorithm [10].

Surgeme Recognition: A surgeme (HMM) is permitted to be followed by any other surgeme during recognition, and the Viterbi algorithm [10] is used to find the sequence $\{\hat{s}_t \in \bigcup \mathcal{S}_\sigma, t = 1, \ldots, T\}$ of HMM states with the highest *a posteriori* likelihood given a test trial $\{x_t\}$. The surgeme sequence $\{\hat{\sigma}_{[i]}, i = 1, 2, \ldots, \hat{k}\}$ and time-marks $[\hat{b}_i, \hat{e}_i]$ are a byproduct of the Viterbi algorithm.

3 Improved Dimensionality Reduction and Modeling

3.1 Linear Discriminant Analysis Based on HMM States

The primary purpose of LDA is to reduce the dimensionality of $\{y_t\}$ without losing information necessary to discriminate between gestures σ_t. Note, however, that each surgeme is modeled by a HMM with several states $s \in \mathcal{S}_\sigma$, each of which models a sub-gesture—called a *dexeme* to connote small dextrous motions. It is natural, therefore, to investigate whether it is better to perform LDA to discriminate between dexemes rather than entire surgemes. An immediate hurdle we face is that the manual segmentation of $\{y_t\}$ is only up to surgemes, and not at the finer resolution of dexemes. But the HMM formalism provides a workaround.

Using the d-dimensional training data $\{x_t\}$ derived from surgeme-level LDA, we first estimate surgeme HMMs as described above, and use the Viterbi algorithm to obtain a *forced alignment* of $\{x_t\}$ with the states of the surgeme HMMs. This results in a dexeme-level segmentation of each surgeme. We use the resulting dexeme *label* $\hat{s}_t$ of each block $[y_{t-p}^{\mathrm{T}} \cdots y_{t-1}^{\mathrm{T}}\, y_t^{\mathrm{T}}\, y_{t+1}^{\mathrm{T}} \cdots y_{t+p}^{\mathrm{T}}]^{\mathrm{T}}$ to compute a new projection matrix $\mathbf{A}$ and use that for all subsequent experiments.

The dexeme-level LDA is better able to preserve information that distinguishes temporal sub-gestures of a single gesture, as well as stylistic variations between samples of the same gesture, as will be demonstrated in Section 3.3.

3.2 Data-Derived HMM Topologies

In the work of [1], and in our initial work here, we used a 3-state left-to-right HMM to model each gesture. However, each gesture has not only temporally distinct sub-gestures—which would be well modeled by states of a left-to-right HMM—but also *contextual* variability in sub-gestures. Some of the latter variability is due to the skill level of the surgeon, some due to the dynamics of a previous or subsequent gesture, while some depends on where in the suturing task (e.g. on the first or fourth stitch) the gesture is being performed. We investigate induction of an optimal HMM topology directly from the data to model such variability.

Formally, we wish to find the topology of a surgeme HMM that maximizes the likelihood (2) of the training data $\{x_t\}$. Finding the optimal HMM topology, however, is computationally intractable: given $n = |\mathcal{S}_\sigma|$, one must find, separately for every n-vertex directed graph, the HMM parameters that maximize (2).

In Speech recognition, HMM topologies are derived for capturing context-dependent (allophonic) variations of phonemes using greedy algorithms. We apply one such algorithm by Varadarajan et al [11], called the modified successive state splitting (SSS) algorithm, to our problem. We begin with a single-state HMM for each surgeme, and iteratively estimate the HMM parameters and increment the number of HMM states via SSS .

Data-derived HMM topologies yield accurate models for surgeme recognition, and also capture sub-gesture patterns indicative of skill, as shown in Section 4.

3.3 Surgeme Recognition and Segmentation Results

We performed surgeme recognition experiments with the training/test partitions described in Section 2.

We first estimated a 1-state HMM per surgeme. In this case, there is no difference between surgeme-level and dexeme-level LDA. The 70% to 74% accuracy for Setup II reported in Table 1(a) may therefore be directly compared with the results of [1], who report accuracies of 64% to 72%.

Next, we estimated a 3-state left-to-right HMM for each surgeme. With surgeme-level LDA, [1] report accuracies of 72% to 77%. In comparison, the dexeme-level LDA provides up to 86% accuracy, as shown in Table 1(b). We also see from Table 1(b) that maximum accuracy is achieved when the number of dimensions d is between 9 and 17 indicating the need for more dimensions to differentiate between the finer grained motions represented by dexemes.

Modeling a surgeme as a temporal sequence of 3 dexemes (left-to-right HMM states) is better than a single-state HMM, but still ad hoc. Determining the HMM topology from data permits modeling both temporally distinct sub-gestures and contextual variability of gestures, as discussed in Section 3.2. Therefore, we use

Table 1. Surgeme Recognition Accuracies with Dexeme-level LDA

(a) A 1-state HMM per Surgeme

LDA d	Setup I	Setup II	Setup III
3	75%	75%	58%
5	81%	72%	69%
7	81%	70%	72%

(b) A 3-state HMM per Surgeme

LDA d	Setup I	Setup II	Setup III
3	79%	70%	73%
5	82%	76%	73%
7	82%	83%	81%
9	82%	**86%**	78%
17	**87%**	83%	**81%**

(c) Data-derived HMM Topology

LDA d	Setup I	Setup II	Setup III
3	69%	67%	64%
4	73%	73%	70%
10	83%	82%	**73%**
15	86%	82%	71%
20	**87%**	**83%**	70%

the SSS algorithm to evolve a 6-state HMM for each gesture. Table 1(c) shows recognition results for the different setups. The recognition accuracies remain high for Setup I and II using data-derived HMMs. The maximum recognition accuracy is obtained when the number of dimensions d is 20, indicating the need for more dimensions needed to differentiate between the larger number of dexemes. We also note that the accuracies drop considerably for Setup III. We conjecture that in addition to expertise-dependent dexemes, the data-derived HMMs may also be modeling user-specific dexemes. This leads to improved recognition when a new trial of a seen user is presented, but also to some *overfitting* to seen users.

The optimal LDA dimension is empirically seen to be proportional to the number of classes: 5 for 1-state HMMs (discriminating 8 surgemes), 9-17 for 3-state HMMs (24 dexemes), and 15-20 for data-derived HMMs (48 dexemes).

4 Surgeme-Level Skills Revealed in Dexeme-Sequences

To illustrate how data-derived HMM topologies encode dexterity information, consider Figure 1, which shows a 5-state HMM derived via the SSS algorithm for surgeme #3 corresponding to the act of "inserting needle through the tissue.". Training samples of surgeme #3 were aligned with this 5-state HMM, and the state-level time marks were used to isolate individual dexemes corresponding to the HMM states a, b, c, d and $e \in \mathcal{S}_3$.

We studied the endoscope video to understand what the segments that align with each dexeme (HMM state) represent, and observed the following.[1]

Dexemes a, b and c: They all constitutes rotating of the right hand patient-side wrist to drive the needle from the entry- to the exit.

Dexeme c versus a and b: All examples that aligned to c were from *novice* surgeons. Examining the videos revealed that c corresponds to a sub-gesture where the novice hesitates/retracts while pushing the needle to the exit point. In most cases, c is followed by a or b, in which the trainee surgeon eventually performs

[1] Video corresponding to these dexemes is available at `www.clsp.jhu.edu/~balakris/MICCAI2009/`

the task (inserting the needle till it exits) correctly. States a and b appear to be indistinguishable, except for some stylistic differences.

Dexeme d: It represents the left arm reaching for the exiting needle. Often, when the left arm is already positioned near the exit point, this gesture is omitted. This explains the transitions from states a and b directly to state e.

Dexeme e: It represents firmly gripping the needle with the left arm.

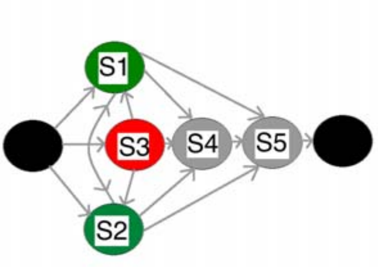

Fig. 1. The Data-derived HMM for $n = 5$ States for Gesture #3

These observations reinforce the claim that *SSS provides a means for automatically inducing meaningful units for modeling dexterous motion.* While not demonstrated here, it may be applied to entire trials, automatically discovering and modeling gestures without requiring any manual labeling!

4.1 Measuring Expertise by Aligning Dexeme-Transcripts

To compare how dissimilar two instances of a surgeme are, we compute an edit distance between their *dexeme transcripts* as described below.

Let $\{x^1_t, t = \hat{b}_i, \ldots, \hat{e}_i\}$ and $\{x^2_t, t = \hat{b}_j, \ldots, \hat{e}_j\}$ denote two *automatically* segmented and labeled realizations of the surgeme σ, i.e. $\hat{\sigma}_{[i]} = \hat{\sigma}_{[j]} = \sigma$. We use the Viterbi alignment of $\{x^1_t\}$ with the states $\mathcal{S}_\sigma$ of the surgeme HMM to obtain the sequence $\{\hat{s}^1_t, t = \hat{b}_i, \ldots, \hat{e}_i\}$, and similarly $\{\hat{s}^2_t, t = \hat{b}_j, \ldots, \hat{e}_j\}$ from $\{x^2_t\}$. We then obtain the sequence of HMM states visited by $\{x^1_t\}$ (resp. $\{x^2_t\}$) by simply compacting each *run* of state labels. In other words, we ignore how many *consecutive* frames are aligned with a state, counting them collectively as one "visit" to the state. Let $\{\hat{s}^1_{[i]}, i = 1, \ldots, \hat{k}^1\}$ and $\{\hat{s}^2_{[j]}, j = 1, \ldots, \hat{k}^2\}$ denote the dexeme transcripts of the two gestures generated in this manner.

We then align $\{\hat{s}^1_{[i]}\}$ and $\{\hat{s}^2_{[j]}\}$ using Levenshtein distance, and each element in the two sequences is marked as matched if it is aligned with the an identical element in the other sequence. Inserted, deleted and (both sides of a pair of) mismatched symbols are marked as mismatched. The *similarity* of the realizations $\hat{\sigma}_{[i]}$ and $\hat{\sigma}_{[j]}$ is defined as the number of matched dexemes divided by $\hat{k}^1 + \hat{k}^2$. A similarity of 1 corresponds to identical dexeme sequences: $\hat{k}^1 = \hat{k}^2$ and $\hat{s}^1_{[i]} = \hat{s}^2_{[i]}$ for each i. Otherwise similarity ranges between 0 and 1.

We calculate the average edit distance between realizations of σ drawn from different expertise levels for the four most frequent gestures: $\sigma = 2, 3, 4$ and 6.

Note from Tables 2(a), 2(b) and 2(c) that some surgemes (e.g. #2 : "positioning the needle at the entry point" or #3 : "inserting the needle through the tissue") show low expert-novice similarity compared to expert-expert, indicating the need for skillful execution. In comparison, surgeme #6 (pulling the suture) in Table 2(d) exhibits significant similarity even between experts and novices.

The correlation between expertise level and edit distance is clearly evident.

Table 2. Dexeme Similarity of Surgemes Performed with Different Skill Levels

(a) Similarities in Surgeme #2

	Expert	Inter.	Novice
Expert	0.65	0.55	0.55
Intermediate	0.55	0.50	0.53
Novice	0.55	0.53	0.46

(b) Similarities in Surgeme #3

	Expert	Inter.	Novice
Expert	0.69	0.60	0.53
Intermediate	0.60	0.51	0.50
Novice	0.53	0.50	0.50

(c) Similarities in Surgeme #4

	Expert	Inter.	Novice
Expert	0.71	0.57	0.54
Intermediate	0.57	0.58	0.58
Novice	0.54	0.58	0.51

(d) Similarities in Surgeme #6

	Expert	Inter.	Novice
Expert	0.74	0.69	0.68
Intermediate	0.69	0.65	0.67
Novice	0.68	0.67	0.61

5 Concluding Remarks and Potential Applications

We have demonstrated the utility of sub-gesture-level LDA in improving dimensionality reduction for HMM-based gesture recognition. We have also shown that data-derived HMMs automatically discover and model skill-specific sub-gestures, leading to a natural metric (dexeme edit distance) for comparing surgical gestures for skill assessment. Since the dexemes are data-derived, such comparison may be feasible even if the manual labeling of surgemes is very coarse grained or absent. Finally, dexeme edit distance based alignment may be transferred to synchronize the surgical *video*, opening up immense possibilities for training.

References

1. Reiley, C., Lin, H., Varadarajan, B., Khudanpur, S., Yuh, D.D., Hager, G.D.: Automatic recognition of surgical motions using statistical modeling for capturing variability. In: MMVR (2008)
2. Shuford, M.: Robotically assisted laparoscopic radical prostatectomy: a brief review of outcomes. Proc. Baylor University Medical Center 20(4), 354–356 (2007)
3. Lenihan Jr., J., Kovanda, C., Seshadri-Kreaden, U.: What is the Learning Curve for Robotic Assisted Gynecologic Surgery? J. Min. Inv. Gyn. 15(5), 589–594 (2008)
4. Martin, J., Regehr, G., Reznick, R., MacRae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (OSATS) for surgical residents. British Journal of Surgery 84(2), 273–278 (1997)
5. Dosis, A., Bello, F., Gillies, D., Undre, S., Aggarwal, R., Darzi, A.: Laparoscopic task recognition using hidden markov models. In: MMVR (2005)
6. Richards, C., Rosen, J., Hannaford, B., Pellegrini, C., Sinanan, M.: Skills evaluation in minimally invasive surgery using force/torque signatures. Surgical Endoscopy 14, 791–798 (2000)
7. Rosen, J., Solazzo, M., Hannaford, B., Sinanan, M.N.: Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden markov model. Computer Aided Surgery 7(1), 49–61 (2002)

8. Lin, H.C., Shafran, I., Murphy, T.E., Okamura, A.M., Yuh, D.D., Hager, G.D.: Automatic detection and segmentation of robot-assisted surgical motions. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 802–810. Springer, Heidelberg (2005)
9. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188 (1936)
10. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
11. Varadarajan, B., Khudanpur, S., Dupoux, E.: Unsupervised learning of acoustic sub-word units. In: Proceedings of ACL 2008: HLT, Short Papers, 165–168 (2008)