Event-based Interpretation of HMM State Sequences for Speech Analysis

K. Sri Rama Murty
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai, India
ksrmurty@gmail.com

B. Yegnanarayana
International Institute of Information Technology
Hyderabad, India
yegna@iiit.ac.in

ABSTRACT

In this paper, we propose a hidden Markov model based approach to capture the effects of core articulatory changes that occur in the speech production mechanism. The changes that are integral to the production of a given sound unit, and must be exercised by all speakers for producing that sound, are considered as *events*. In this approach, the events are interpreted from a suitable subset of the state sequences of a hidden Markov model. An event is associated with a probability value, and a label to represent the significance and the nature of the event, respectively. Using this approach, a given sound unit can be represented as a sequence of events. The consistency of the sequence of events across different speakers is demonstrated by performing digit recognition experiments. It is to be noted that the objective of these experimental studies is not to develop a recognition system, but to provide an event-based interpretation for the speech signal through a subset of state sequences of the hidden Markov model.

Keywords

Event, hidden Markov model, state transition sequence, event probability sequence, isolated digit recognition.

1. INTRODUCTION

Speech can be considered as a sequence of events, where an event can be interpreted as change in some characteristics of the speech production reflected in the speech signal. The events occurring in the speech production process can be viewed at various levels such as signal level, production level, acoustic level, phonetic level, sound unit level, suprasegmental level, speaker level and language level. In this work, we propose a hidden Markov model based approach to detect the commonly occurring events across the speakers, at production and acoustic levels, while uttering a given sound unit.

At production level, speech may be characterized in terms of production features such as voicing, aspiration, frication and burst. Onset of any of these features and change from one feature to the other may be treated as events at the production level. The characteristics of the vocal tract (acoustic) system depends on positioning of various articulators, which in turn decides the type of speech sound produced. The changes in the positioning of articulators may be treated as events at the acoustic level. For instance, during the production of bilabial sounds, opening of lips from initial closure is an event. Though some of these events are speaker-specific, there exist certain core events that should occur commonly across speakers while producing a given sound unit. For example, during the production of consonant-vowel $/ba/$ one has to necessarily close lips. At the perception level also, human beings do not convert a speech signal continuously into subwords or words as automatic speech recognition systems attempt to do. Instead, human beings seem to detect acoustic and auditory evidences, weigh them and combine them to form cognitive hypothesis, and then validate the hypothesis until consistent decisions are reached. This process has been successfully demonstrated in spectrogram reading by experts trained in acoustic-phonetics [1]. Hence, speech production and perception mechanisms of human beings function mainly on certain important events that are present in the speech signal. Based on these two aspects, the efforts for development of an automatic speech recognition system have resulted in two different directions of speech research.

Most of the speech recognition systems concentrate on recognizing what a human being actually perceives from a speech signal. Since the human perception mechanism is less understood, the research in this area is mainly concerned about building statistical models for the subword units like phoneme or syllable, and then using syntactic and semantic constraints to recognize words and sentences. In spite of reasonable success, these methods are often criticized for having little relation to the actual human way of speech production/perception [2]. On the other hand, the motivation for structural representation of the speech comes from the theory of articulatory phonology [3]. In this approach, the vocal tract activity during speech production is decomposed into discrete, recombinable atomic units. Compared to traditional approaches based on phone models and syllable models, structural approach is more concrete physiologically, and offers a compact means of representing the speech signal [4]. A major drawback is that the speech signal collected through a microphone alone is not enough to detect these gestures. The present methods for direct articulatory measurements are strongly dependent on X-ray techniques, such as cineradiography of human head during speech production [5]. Accurate estimates of articulatory measurements may not be required for applications like speech recognition, because of the inter speaker variations in the articulatory measurements. Instead, the articulatory information expressed in terms of highly quantized abstract classes (voicing, lip rounding, nasality, etc.) is widely used in speech recognition systems.

Several approaches have been proposed to bridge the gap between acoustic and articulatory modeling techniques in the context of speech recognition. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features is proposed in [6]. The overlapping articulatory feature model aims at constructing a multidimensional hidden Markov model (HMM), whose states can be made to directly correspond to the symbolically coded, phonologically contrastive articulatory structure responsible for generating acoustic observations from the states [2]. In [7], Kirchhooff et al., demonstrated a system based on artificial neural networks to estimate gross articulatory features (voicing, lip rounding, place and manner of articulation) from acoustic features. In this paper, we propose a method to explore a suitable subset of state sequences of HMM that may bring out some characteristics of events related to the speech production process. This paper is organized as follows. A brief overview of the traditional HMM based approaches and a method to detect the events from the state sequences of HMM is presented in Sec. 2. Isolated digit recognition experiments conducted to evaluate the consistency of the events obtained by the proposed method are reported in Sec. 3. In Sec. 4, an efficient way of detecting the events using the partial observation sequence is described. The scope of the work and some of the possible extensions are discussed in Sec. 5.

2. EVENT DETECTION USING HMM

The speech production mechanism is guided by an inherent system which constrains the articulatory movements during the production of sound units. As a result, we cannot produce a given sound unit with an arbitrary sequence of articulatory positions. There exist some core sequence of articulatory movements (events) that should commonly occur across speakers while pronouncing a given sound unit. We cannot observe these events in the samples of speech signals because the fluctuations in the raw data make it difficult to interpret any change as an event. At the feature level too, where a feature vector is derived from a block of samples, it is difficult to distinguish between events and nonevents. Information present in the speech signal is mainly due to the sequence of frames rather than due to any particular frame in isolation. Although, the spectral content of the speech signal may include frequencies up to several thousand Hertz, the articulatory configuration (the vocal tract shape, velum, tongue, lip movements etc.) may not undergo dramatic changes more than ten times per second on the average. It is reasonable to assume that there exist stable states in the speech production process, and gradual transitions occur between these stable states. Under these conditions, the HMM is a better choice to capture the unknown (hidden) state sequence from the observed sequence of feature vectors.

The HMM can be considered as a Markov model in which the observation is a probabilistic function of the state [8]. The HMM can be described by the parameter set $\lambda = (\Pi, A, B)$, where Π defines the initial state probability, $A = [a_{ij}]$, denotes the probability of making a transition from state i to state j, B defines the distribution of feature vectors in each state. The parameter set λ can be estimated using Baum-Welch algorithm [8]. Once the parameters of the HMM are evaluated using the training data, the likelihood $P(\mathbf{O}/\lambda)$ with which a test observation sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$ is obtained from a given model λ is computed as either sum or maximum over all possible state sequences. This is given below:

- Sum over all the possible state sequences,

$$P(\mathbf{O}/\lambda) = \sum_{q_1 q_2 \cdots q_T} P(q_1 q_2 \ldots q_T \mathbf{o}_1 \mathbf{o}_2 \ldots \mathbf{o}_T/\lambda)$$

- Maximum over all the possible state sequences,

$$P_{\max}(\mathbf{O}/\lambda) = \max_{q_1 q_2 \cdots q_T} P(q_1 q_2 \ldots q_T \mathbf{o}_1 \mathbf{o}_2 \ldots \mathbf{o}_T/\lambda)$$

At a gross level, the computation of $P(\mathbf{O}/\lambda)$ and $P_{\max}(\mathbf{O}/\lambda)$ corresponds to the two extreme cases. The computation of $P(\mathbf{O}/\lambda)$ incorporates all the possible N^T state sequences, whereas the computation of $P_{\max}(\mathbf{O}/\lambda)$ considers only the optimal state sequence (one among N^T possibilities). It is always more meaningful to consider a subset of these N^T state sequences (suboptimal state sequences), and observe any commonality among them [9]. This paper attempts to explore any (hidden) sequence of events that may be present in the sequence of states, and not directly from the samples of the speech signal.

2.1 Exploring events in the sequence of states

As described earlier, we assume the speech production process to stay in somewhat stable states and make gradual transitions between these stable states. Our main goal is to capture the nature and the time instants of these transitions, which we call as *events*. The key idea in capturing the events is that, the gross nature of the event sequences obtained for a given word is expected to be similar across speakers, though there may be some missing and spurious events. We define a variable $\eta_t^p(i,j)$ as follow:

$$\begin{aligned}\eta_t^p(i,j) = \quad & P(q_{t-p} = i, q_{t-p+1} = i, \ldots, q_t = i, q_{t+1} = j, \\ & q_{t+2} = j, \ldots, q_{t+p+1} = j/\mathbf{O}, \lambda) \qquad (1)\end{aligned}$$

Here, the p in the superscript refers to consideration of p additional frames before and after the transition. We refer to p as *support* given for the stable state before and after the occurrence of the event at time t. The term $\eta_t^p(i,j)$ can be written as

$$\begin{aligned}\eta_t^p(i,j) \quad = \quad & \alpha_{t-p}(i) a_{ii}^p b_i(\mathbf{o}_{t-p+1}) b_i(\mathbf{o}_{t-p+2}) \ldots \\ & b_i(\mathbf{o}_t) a_{ij} b_j(\mathbf{o}_{t+1}) b_j(\mathbf{o}_{t+2}) \ldots \\ & b_j(\mathbf{o}_{t+p+1}) a_{jj}^p \beta_{t+p+1}(j)/P(\mathbf{O}/\lambda) \qquad (2)\end{aligned}$$

where α and β are the forward and backward variables respectively [8]. We define one more variable e_t^p, similar to that of Viterbi maximization in HMM, as

$$e_t^p(k,l) = \max_{i,j} \eta_t^p(i,j), \qquad i \neq j, \qquad (3)$$

where

$$(k,l) = \arg\max_{i,j} \eta_t^p(i,j), \qquad i \neq j. \qquad (4)$$

Here, $e_t^p(k,l)$ represents the probability with which there can be transition between the stable states k to l, with a support of p frames for the stable states, at the time instant t. We call $e_t^p(k,l)$ as the event probability. A large value of e_t^p indicates the presence of an event, and the corresponding (k,l) indicates the state transition responsible for the event. The nature of the event is specified by (k,l), and the intensity of the event is specified by the value of e_t^p.

During the speech production process, gradual transitions occur between the stable articulatory positions. Hence we cannot expect any abrupt changes in the speech signal. Instead, the changes that occur in the speech signal are continuum in nature. Hence we hypothesize an event at every frame in the speech signal, and evaluate the probability of the event e_t^p and the nature of the event (k,l) that indicates transition from state k to state l. Therefore, at a higher level of abstraction, a given speech signal is represented with an event probability sequences and a state transition sequence. The event probability value is used to decide presence or absence of the event.

In this method of event detection, we have assumed an ergodic HMM, where the transition from present state to any state is allowed with a nonzero probability. Though the ergodic HMM is popular among text-independent speaker recognition studies [10], it is not commonly used in speech recognition studies. The ergodic HMM is used in text-independent speaker recognition because each state is expected to capture a broad phonetic category, and any phonetic category can follow any other depending on the sentence uttered. On the other hand in speech recognition where the models are typically built on subword units like syllable/phoneme, whose signal properties change over time, the left-right model is more suitable since the state index increments with the time [8]. However, the ergodic model is more suitable for the proposed event detection method for the following reasons. If a left to right model is used instead, the state transition probabilities $[a_{ij}]$ for $j < i$ are zero, and the only possible transition from state i is to state $i + 1$. Hence, the state transition sequence obtained consists of only the adjacent transitions $(i, i + 1)$ irrespective of the sound unit, which may be inadequate to capture the salient events in the speech utterance.

3. EXPERIMENTAL EVALUATION OF EVENT-BASED ANALYSIS

The TI isolated digit database was used to evaluate the consistency among the events obtained across different speakers for a given word. The vocabulary consists of 11 words, namely, the 10 digits and “oh”. In the experimental studies, we have considered only the 10 digits and omitted “oh”. In the present studies, we have used the utterances from only the male speakers in the database. There are 111 male speakers, divided into training set (55 speakers) and test set (56 speakers). Each person spoke 20 utterances, consisting of 2 tokens from each of the 10 digits in isolation. Therefore, the training set consists 110 repetitions per digit and the testing set consists of 112 repetitions per digit. All the digit utterances were sampled at 8 kHz. The first 8 MFCCs other than the zeroth value (average of log-spectral values) and their derivatives are used as features to represent the information in the speech signal. A five state ergodic HMM with two mixtures per state is used to model the sequence of feature vectors extracted from the speech signal. One HMM is trained for each digit using the examples of that digit.

Once the HMM is trained, for every digit utterance in the training set, the Event Probability Sequence (EPS) and the State Transition Sequence (STS) are obtained as explained in Sec. 2. The value of p is arrived at empirically after some preliminary experimental studies. A small value of p may produce several spurious peaks. On the other hand, as p is increased, fewer probabilities are used in the summation of (2), and hence the results may become unreliable. A value of $p = 7$ is used throughout the remaining part of experimental studies. The event sequences obtained for the training utterances are used as reference during testing phase. The EPSs and the STSs for the digit *one* uttered by five speakers, when tested against the HMM developed for the digit *one*, are shown in Fig. 1. The STSs are consistent across the five utterances spoken by different speakers, except for a few missing and spurious transitions. The EPSs for the digits *one*, *two*, *three*, *four* and *five* uttered by same speaker, when tested against the HMM developed for digit *one*, are shown in Fig. 2. The STS obtained for the digit *one* is different from the STS obtained for the other digits. Therefore, the STS is a representative of a digit, and hence can be used for digit recognition.

3.1 Matching the sequence of state transitions

In this study, Levenshtein distance [11] was used as a measure of similarity between two event sequences. Levenshtein Distance (LD) or edit distance between two strings is defined as minimum number of point mutations needed to transform one string to another string, where a point mutation can be either insertion, deletion or substitution. Digit recognition experiments were conducted using the LD as measure of similarity between the reference and test STSs. The testing process of event based digit recognition is shown in the form of block diagram in Fig. 3. The STS obtained for the test utterance is matched against each of the reference STS. The accumulated score over all the reference utterances is used to make the decision. The k-best recognition performance of the LD based transition sequence matching for isolated digit recognition is given in Table 1. It has been observed that $nine$ was recognized as either one, $three$ or $five$ in 50% of the cases, and two was recognized as $zero$ in 25% of the cases. From Table 1, it can be seen that though the 1-best recognition performance is only 86%, the 2-best performance is 96%, showing that there is a consistency among the STSs that are obtained across different speakers.

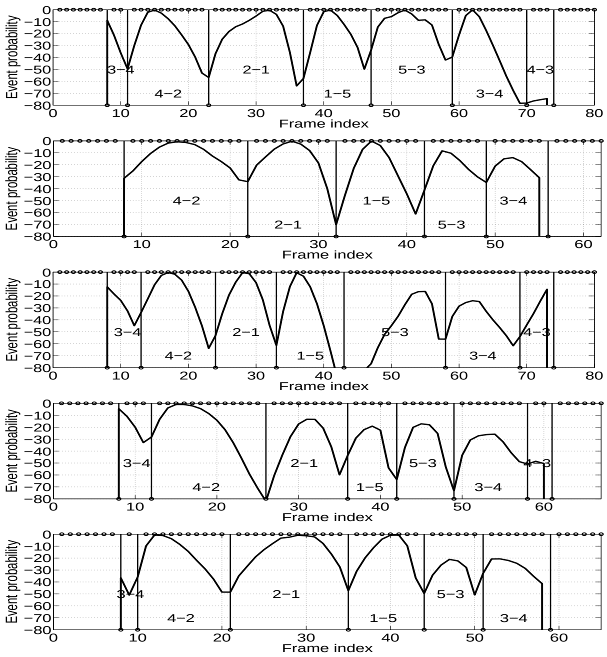

Figure 1: *Event probability sequence for the digit one uttered by five speakers, when tested against model for digit one. The probability value is plotted on logarithmic scale. The state transition $(k - l)$ responsible for the event is also indicated. The vertical lines indicate the span of a particular state transition.*

Though the STSs are consistent among different utterances of the same digit and inconsistent among utterances of different digits, the values of the event probabilities are at the same level in both the cases (see Fig. 1 and Fig. 2). Hence the EPSs do not provide any complementary information to improve the performance of the recognition system. This is because of our definition of $\eta_t^p(i, j)$ given by (2), where we divide the probability of having a transition between two stable states i and j at time instant t by $P(\mathbf{O}/\lambda)$, the probability of the observation sequence $\mathbf{O}$ given the model λ. If the test observation sequence is actually generated by the model, then both the numerator and the denominator of (2) are high, resulting in a value around one. If the test observation sequence is not generated by the model, then both the numerator and the denominator are small, resulting again in a value around one. If the probability of having a transition is not normalized by $P(\mathbf{O}/\lambda)$, the computation of $\eta_t^p(i, j)$ depends on the length of the observation sequence. To overcome this computational limitation, we propose an event detection method based on the partial observation sequence.

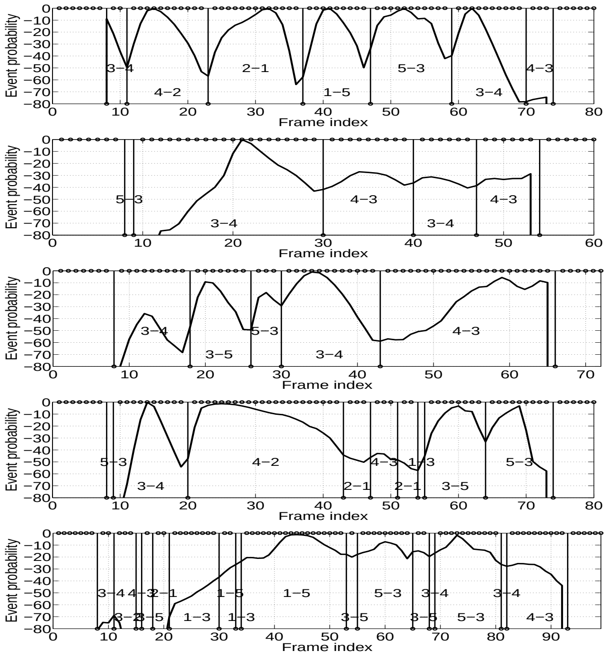

Figure 2: *Event probability sequences for the digits* one, two, three, four *and* five *uttered by the same speaker, when tested against model for digit* one.

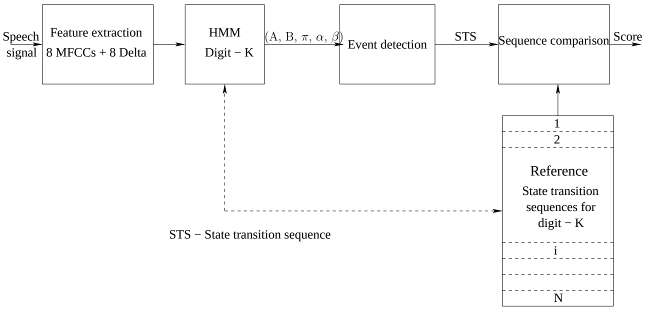

Figure 3: Block diagram for testing process of event based digit recognition

4. EXPLORING EVENTS IN PARTIAL OBSERVATION SEQUENCE

From the previous discussion on the event-based approach, it can be understood that the events are localized in time. In other words, an event occurring at a time instant t is mainly influenced by the symbols in its immediate neighborhood rather than the symbols away from it. Hence, it is better to explore the events within a partial observation sequence rather than considering the entire observation sequence. In this approach, we compute $\eta_t^p(i,j)$ over a partial sequence of $2(p+r)$ frames, where p is the number of frames supporting the stable state, and r is the number of frames allowed on either side of the stable states which can be emitted by any state in the model. Hence $\eta_t^p(i,j)$ is modified as

$$\begin{aligned}\eta_t^{pr}(i,j) \quad &= \quad P(q_1,\ldots,q_r,q_{r+1}=i,\ldots,q_{r+p}=i,\\ &\quad\quad q_{r+p+1}=j,\ldots,q_{r+2p}=j,q_{r+2p+1},\ldots,\\ &\quad\quad q_{2(r+p)},\mathbf{o}_1,\mathbf{o}_2,\ldots,\mathbf{o}_{2(r+p)}/\lambda) \qquad (5)\end{aligned}$$

The value of $\eta_t^{pr}(i,j)$ can be evaluated in a computationally efficient way by using the locally defined forward variable and backward variable on the partial observation sequence $\mathbf{O} = (\mathbf{o}_1,\mathbf{o}_2,\ldots,\mathbf{o}_{2(p+r)})$. Since a fixed number of observation symbols is involved in the computation of the $\eta_t^{pr}(i,j)$, it does not depend on the length of the observation symbol sequence. Moreover, since we are not dividing by $P(\mathbf{O}/\lambda)$ term in the computation of $\eta_t^{pr}(i,j)$, a clear evidence can be observed, for genuine cases, from the EPSs. Therefore, the EPS obtained from this computation can be used along with the STS to further improve the recognition performance. By allowing self transition in the computation of e_t^{pr}, we can segment the speech signal into stable and transition regions. Fig. 4 shows the EPSs computed from partial observation sequences for digit *one* uttered by five speakers when tested against the model *one*. The regions marked by self transitions $(i-i)$ correspond to the stable regions, and the regions marked by the transition between two different states ($i-j$ and $i \neq j$) correspond to transition regions. The EPSs computed from partial observation sequence for the digits *one*, *two*, *three*, *four* and *five* when tested against the model *one*, are shown in Fig. 5. In Fig. 4, the event probability values are high and the STSs are similar across the two repetitions of the digit *one* by different speakers. Hence, the event probability values can be used along with the STS for recognition purpose.

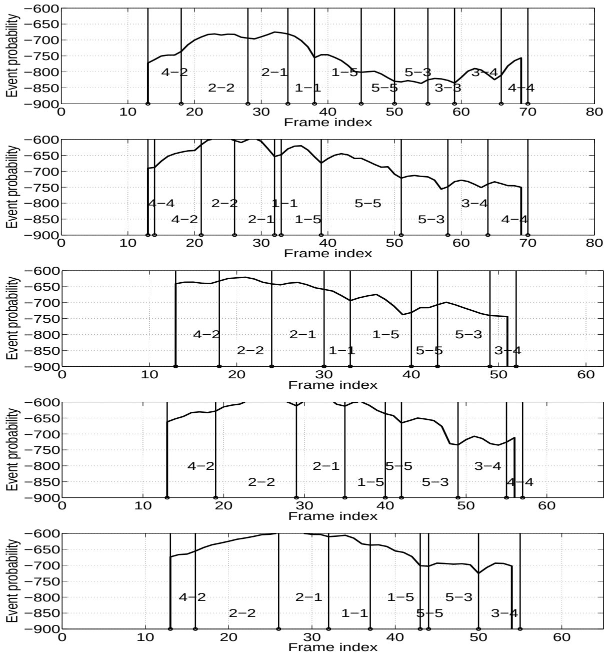

Figure 4: *Event probability sequences computed from partial observation sequence for the digit* one *uttered by five speakers when tested against model for digit* one.

4.1 Event probability sequences for digit recognition

Digit recognition studies are performed using the EPSs of transition regions. Since we are interested in the events (core changes), we have not considered the stable regions. With in a transition region, we have considered only the point where the event probability value is maximum. The product of the maximum event probability values of all the transition regions in a digit is considered as the confidence score. In this method of scoring, an average recognition performance of 94% was obtained. The k-best performance

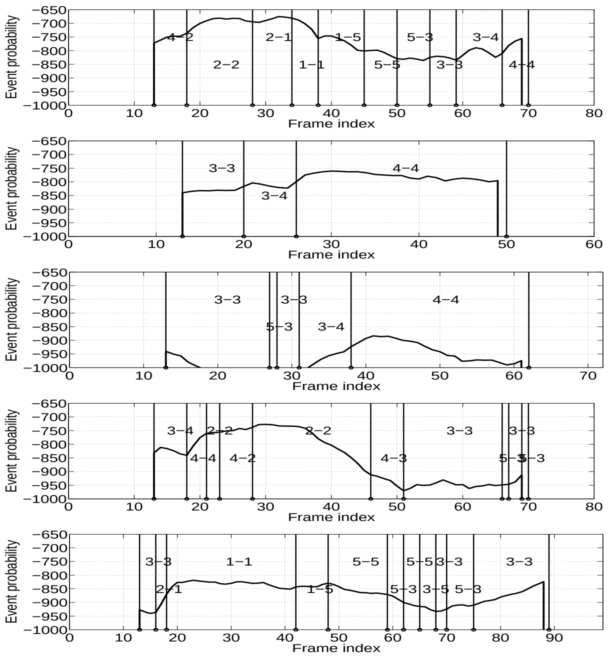

Figure 5: *Event probability sequences computed from partial observation sequence for the digits* one, two, three, four *and* five *when tested against model for digit*one.

analysis based on the EPSs is given in the Table 1. In this case it has been observed that $nine$ was recognized as $five$ in 19% of the times, and was not confused with one. The recognition performance can be improved by combining the evidences from EPSs and the STSs.

The confidence scores C_s and C_p obtained using the STSs and the EPSs, respectively, are combined using a linear weighted sum, given by $C_c = aC_s + (1 - a)C_p$. The parameter $a(< 1)$ governs the weighting given the individual scores. Here, a value of $a = 0.5$ is used to combine the evidences. The k-best performance analysis is given in the Table 1. By combining the evidences from the STSs and the EPSs, an average performance of 97% was obtained, which is significantly better than both of the individual systems. This shows that the STS along with the EPS provide unique signature for a given sound unit.

Table 1: *k-best performance digit recognition for state transition sequence (STS), event probability sequence (EPS) and combined system.*

Digit	STS		EPS		Combined	
	k=1	k=2	k=1	k=2	k=1	k=2
1	98	100	95	100	99	100
2	66	92	95	98	96	98
3	97	100	98	100	99	100
4	90	99	99	100	99	100
5	77	96	100	100	100	100
6	97	99	87	97	99	100
7	88	99	93	98	96	99
8	93	96	97	98	97	98
9	47	82	79	100	82	100
0	99	100	98	99	100	100

5. SUMMARY AND CONCLUSION

This paper describes a probabilistic approach to examine a subset of the state sequences in the HMM to determine if they can be interpreted in terms of sequence of some events. The intention is to eventually relate these sequences of events to some key articulatory movements in the speech production process. The sequence of events is detected by computing the probability of making a transition between two hidden states with a support of p frames on either side of the transition. Experimental results on isolated digit recognition indicate that the event sequences (STS along with the corresponding EPS) represent a given sound unit at a higher level of abstraction. Since the event sequences appear to be consistent across different speakers, it will be interesting to provide an acoustic-phonetic or articulatory description of the events by analyzing speech signal in the regions around the events.

6. REFERENCES

[1] V. W. Zue, “Acoustic-phonetic knowledge representation: Implications from spectrogram reading experiments,” in *NATO ASI on speech recognition*, Bonas, France, 1981.

[2] L. Deng, G. Ramsay, and D. Sun, “Production models as a structural basis for automatic speech recognition,” *Speech Communication*, vol. 22, no. 2-3, pp. 93–111, 1997.

[3] C. Browman and L. Goldstein, “Articulatory phonology: An overview,” *Phonetica*, vol. 49, pp. 155–180, 1992.

[4] G. N. Clements, “Phonological primes: Features or gestures?,” *Phonetica*, vol. 49, pp. 181–193, 1992.

[5] J. Papcun, T. R.Hochberg, F. Thomas, J. Larouche, J. Zacks, and S. Levy, “Inferring articulation and recognizing gestures from acoustics with a neural network training on X-ray microbeam data,” *J. Acoust. Soc. Amer.*, vol. 92, no. 2, pp. 688–700, 1992.

[6] Li Deng and D. Sun, “A statistical approach to automatice speech recognition using the atomic speech units constructed from overlapping articulatory features,” *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2702–2719, 1994.

[7] Katrin Kirchhoff, Gernot A. Fink, and Gerhard Sagerer, “Combining acoustic and articulatory feature information for robust speech recognition,” *Speech Communication*, vol. 37, no. 3-4, pp. 303–319, Jul. 2002.

[8] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[9] Naresh P. Cuntoor, B. Yegnanarayana, and R. Chellappa, “Interpretation of state sequences in HMM for activity representation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, Mar. 2005, pp. 709–712.

[10] Tomoko Matsui and Sadaoki Furui, “Comparision of text-independent speaker recognition methods using VQ-distortion and discrete/continous HMMs,” *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, pp. 456–459, Jul 1994.

[11] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” *Doklady Akademii Nauk SSSR*, vol. 163, no. 4, pp. 845–848, 1965.