

REGENERATION TIME, MEMORY, AND ENTROPY IN SPEECH RECONSTRUCTION

Dr. Michael D. Moore, Ph. D.

AetherMachines Inc., Alplaus, NY 12008, USA www.aethermachines.com

ABSTRACT

State based speech reconstruction is a relatively new application for Markov models such the HMM and HSMM (Hidden Semi – Markov Model). Generalized HSMMs (GHSMMs) contain additional memory that creates a non – stationary, time varying topology. This topology implements longer *regeneration times* (times at which a Markov chain becomes independent of its extended past) that increase reconstruction accuracy. The statistics of these longer regeneration times provides some very interesting insights into the roles of regeneration time, memory, and entropy in speech reconstruction using Markov models. Key words : HMM, HSMM, GHSMM, regeneration time, Zipf’s Law.

1. INTRODUCTION

Speech reconstruction has typically been attempted at the level of individual speech features. This approach, akin to speech enhancement, attempts to blend reconstruction with recognition by improving damaged features of individual speech frames. Our research using GHSMMs takes a markedly different approach for reconstructing English in an aviation vocabulary. Damaged frames are identified first using a spectral noise model and tagged as ‘unknown’. This unknown tag is equally likely in all Markov model states during reconstruction. A generic recognizer is used to classify the undamaged frames. Markov models (including the GHSMM) are then used to statistically reconstruct the state sequence using known (white) frames to determine the unknown (gray) frames. A database of phonemes is kept for a specific speaker and the corrected state sequence is re – spoken in the proper voice to complete reconstruction (Figure 1) [2,3].

2. HMMS, HSMMS, AND GHSMMS

HMMs are well known stochastic models containing transition probabilities A , observation probabilities B , and states $s_i \in s$. A widely known drawback of HMMs is the

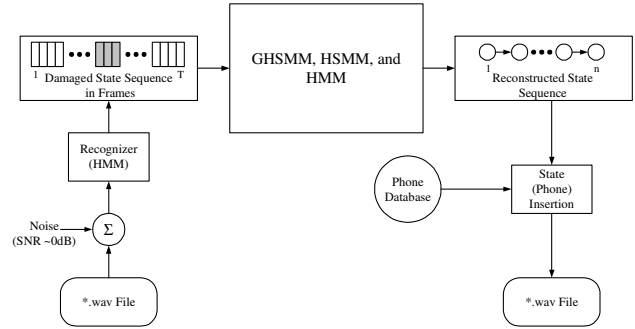


Figure 1. The Reconstruction System

geometric state duration modeling implicit in the state self transitions contained on the diagonal of the A matrix. HSMMs incorporate arbitrary state duration distributions in an additional matrix D . States $s_i \in s$ in HSMMs persist by virtue of these distributions, so the diagonal elements (self transition) in the A matrix of an HSMM are all zero. The distributions D are thus simply regeneration times on the order of one speech state [2,3].

2.1. The GHSMM

Generalized HSMMs (GHSMMs) incorporate additional memory in the transition matrix A . This memory results in a non – stationary topology and implements an arbitrary regeneration time (Figure 2). The elements a_{ij} of the A matrix become a function of transition time of state pair $s_i s_j$ and membership this pair in a longer regeneration time (state sequence or n – Window

$\omega = s_1 s_2 \dots s_{n-1} s_n$, Figure 3) [2,3] :

$$\tilde{a}_{ij} = a_{ij} P((s_i s_j) | t) P((s_i s_j) \in \omega = (s_1 s_2 \dots s_{n-1} s_n)) \quad (1)$$

2.2. Transition time probability

The transition time probability $P((s_i s_j) | t)$ is generated from a longest expected regeneration time T by a simple uniform probability (here described with step functions) :

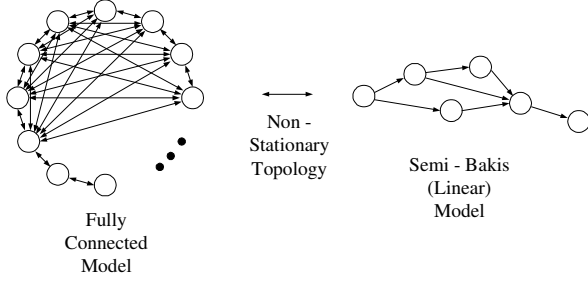


Figure 2. Non – Stationary Topology

$$P((s_i s_j) | t) = \frac{u[t - (t_{ij} - \varepsilon)] - u[t - (t_{ij} + \varepsilon + 1)]}{(t_{ij} + \varepsilon + 1) - (t_{ij} - \varepsilon)}$$

$$0 \leq \varepsilon \leq T,$$

$$P(1 \leq t \leq T_{max}) = 1.0 \quad (t_{ij} - \varepsilon) = \max(t_{ij} - \varepsilon, 0),$$

$$(t_{ij} + \varepsilon + 1) = \min(t_{ij} + \varepsilon + 1, T) \quad (2)$$

In practice, $P((s_i s_j) | t) \geq \eta$, $0 < \eta \ll 1 \forall t$, to always permit some temporal misalignment [2,3].

2.3. Random field probability

$P((s_i s_j) \in \omega = (s_1 s_2 \dots s_{n-1} s_n))$ is the probability of membership of a state pair (or equivalently a state transition) in a sequence of n states ω as given by a random field. The use of a random field has its origins in Zipf's law (the so – called principle of least effort), and roughly assumes that deviation from accepted (meaningful) state sequences is exponentially detrimental to the conveyed meaning [2,3,4]. A simple cellular automaton calculates candidate regeneration times $\omega \in w^*$ from a secondary observation produced by the known states. This secondary observation is simply those state sequences that the known state transitions $s_x s_y$ have participated in. Coupling function $K(\varepsilon, \xi) = \xi(1 - \frac{2\varepsilon + 1}{T})$ is used to combine field strength ξ with uncertainty ε from (2) [1,2] :

$$P(\omega = (s_1 s_2 \dots s_{n-1} s_n)) = \frac{e^{-\xi(1 - \frac{2\varepsilon + 1}{T}) \sum_T \min_k |w^* - w_{ij}(k)|}}{Z}$$

$$\propto P((s_1 s_2) \in \omega) \dots P((s_{n-1} s_n) \in \omega) \quad (3)$$

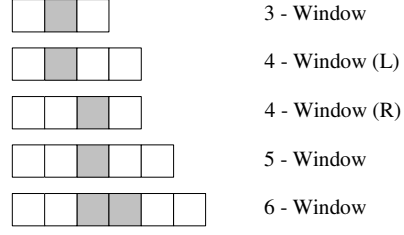


Figure 3. Various Size n – Windows ω (Regeneration Times)

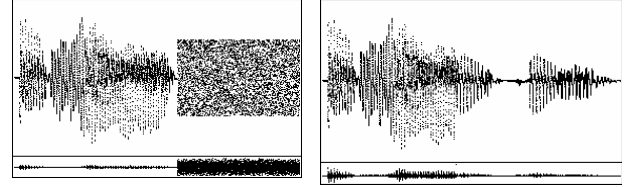


Figure 4. The Word 'American', Damaged (Left) and Reconstructed (Right)

Partition function Z is calculated in a straightforward manner to normalize probabilities and (3) is then interleaved in a HSMM Viterbi algorithm using the independence of the sum. The most likely state sequence(s) are then calculated to perform reconstruction. In the sum contained in (3), the random field potential function is formed from a Manhattan distance between state sequences $\omega = w_{ij}(k)$ (regeneration times) when those sequences are arranged in a canonical sorted order [2,3] :

$$V(\omega_{ij}) = \min_k |w^* - w_{ij}(k)| \quad (4)$$

Additional memory $\omega = w_{ij}(k)$ contains the canonical sorted location of the longer regeneration times ω that transition $s_i s_j$ has participated in. This memory, in conjunction with the transition time memory implied in (2), create the non – stationary topology of Figure 2 during repeated Viterbi passes through the GHSMM lattice [2,3].

3. RECONSTRUCTION RESULTS

Because of the power of longer regeneration times captured in (1), the GHSMM outperforms the HMM and HSMM when reconstructing n – Windows and entire words. Figure 4 illustrates example damage to one of several hundred utterances when reconstructing with the system of Figure 1. The SNR in Figure 4 (and for all

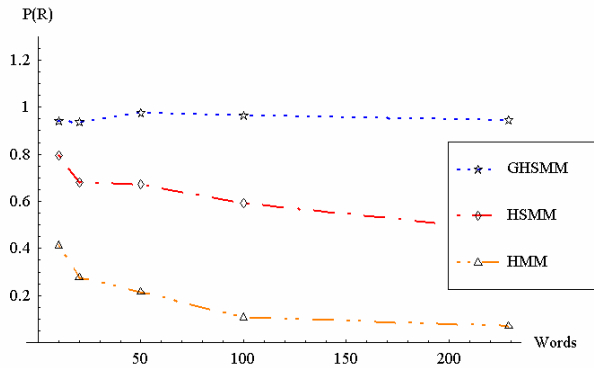


Figure 5. Reconstruction Results for n – Windows

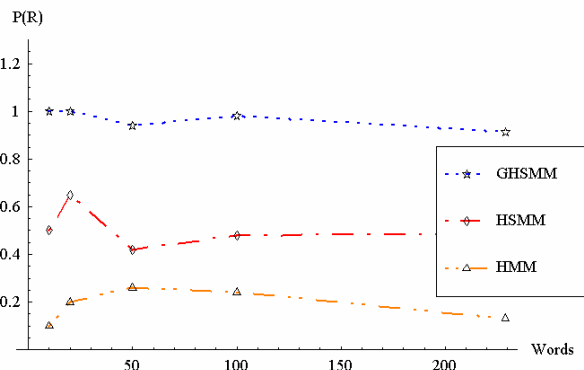


Figure 6. Reconstruction Results for Words

reconstructions) is approximately 0dB. All damage is randomly placed for random durations. A reconstruction is correct if it contains the correct speech states in time order; otherwise it is incorrect. The reconstruction success rate $P(R)$ is the ratio of correct reconstructions to total attempts [2,3].

Figures 5 and 6 illustrate the reconstruction advantage of the GHSM for n – Windows (pieces of words) and entire words respectively. Both Figure 5 and 6 plot reconstruction rate vs. the number of words in an increasingly large vocabulary [2,3].

3. REGENERATION TIMES, MEMORY, AND ENTROPY

In larger corpora such as the DARPA TIMIT, regeneration times expressed as n – Windows have the Zipf – like ranking behavior shown in Figure 7 [2,3]. This behavior consists of a relatively few ‘frequent’ state sequences followed by a horde of ‘equi – probability’ sequences, those that appear just once or twice in the training corpus and/or general use [2,3].

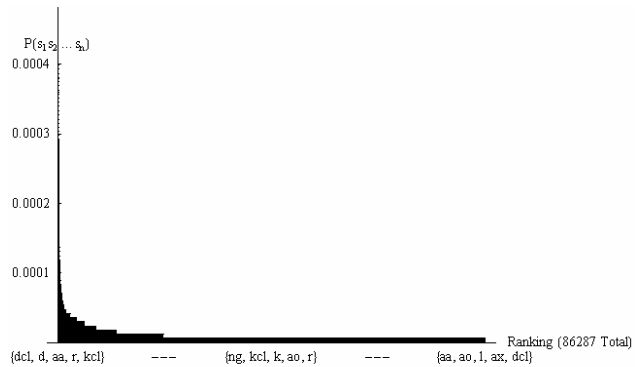


Figure 7. Zipf – Like Ranking of 5 – Windows in DARPA TIMIT

The HSMM, by virtue of the arbitrary state durations in matrix D , has a regeneration time on the order of a state and cannot use the longer state relationships implied by Figure 7. The HSMM can thus reconstruct contiguous unknown frames quite well, as long as the duration of the damage is less than an average state. The GHSM contains much longer regeneration times, and as such can leverage the longer time behavior implied in Figure 7 [2,3].

As n – Window size increases, plots like Figure 7 become flatter and less probability is contained in the exponential drop at the beginning of the plot. More probability is contained in the equi – probability ‘dust’, until the great majority of state sequences (regeneration times) are all equally likely [1]. It is interesting to note that in this case the common assumption of equal priors for specific Markov models (as when a Markov model is kept for each word in recognizers) becomes completely valid.

Another interesting observation is that although most very long state sequences become equally likely, and hence the ‘statistics’ of decision making is obviated (there is really no statistical advantage to choosing any state sequence), the number and length of the state sequences remembered makes statistics ‘unnecessary’. This is because longer regeneration times associated with longer n – Windows ‘span the space’ of possible longer sequences so completely that highly discriminative statistics are no longer needed in very structured vocabularies such as aviation communications. This is of course only applicable when the training of the GHSM has sufficient data to provide large numbers of longer n – Windows (regeneration times) [2,3].

Longer regeneration times also have compelling possibilities for capturing longer statistical behaviors,

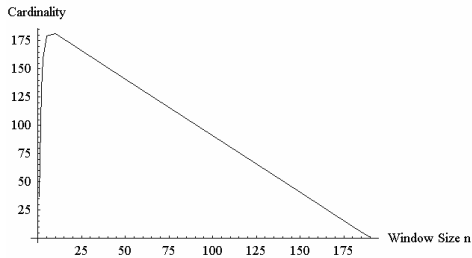


Figure 8. ATIS Example Cardinality of n – Window Sets

whether through the curved ‘statistical’ or flat ‘memory’ regions of Figure 7. Such behaviors can be illustrated using an ATIS/AWOS (automated weather) aviation example :

“Schenectady County Airport – automated weather observation – one two one two weather – wind zero five zero at one two – peak gusts one seven – visibility one zero – clear below one two thousand – temperature one five Celsius – dewpoint one four – altimeter two niner niner two”

When recognized as states (phones), this sentence becomes a sequence 190 states in length, comprised of 37 individual phones. The statistics of one, two, three five, ten, 50, 150, 180, 188, 189, and 190 – Windows in this sentence illustrate the ‘statistics and memory’ associated with regeneration times. Figure 8 shows the cardinality of the n – Window sets with increasing window size (regeneration time) n and Figure 9 the entropy of each set expressed in bits against the same.

For example, the regeneration time of a 50 – Window can capture the statistics of the relationship of locally prevailing wind speeds and wind directions, daily temperature, dewpoint and pressure variations, and perhaps even visibility, temperature, dewpoint, and peak wind gust relationships (as a predictor of convective activity in the vicinity).

Conversely, using the regeneration time of the 3 – Window (the length of the word five, for example) could be disastrous. If the most likely windspeed reconstructed is unlikely for the wind direction, a dangerous runway could be chosen. If the most likely visibility is reconstructed without factoring in ceiling, temperature and dewpoint, an ‘illegal’ VFR (visual flight rules) takeoff could be attempted !

A final intriguing observation is that both set cardinality and entropy initially increase with increasing regeneration time. After a certain extremum, entropy decreases nonlinearly with set cardinality.

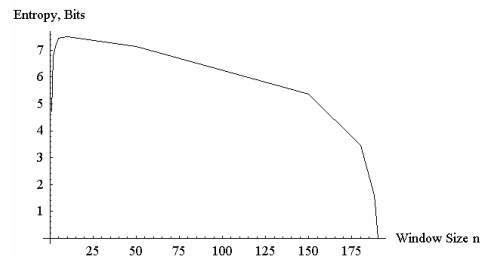


Figure 9. ATIS Example Entropy of n – Window Sets

At an upper regeneration time (n – Window size), the entropy of the sets of 1 – Windows (single states) and n – Windows becomes equal again.

4. CONCLUSION

Regeneration times allow for quite accurate state sequence reconstructions for English speech. The Zipf – like behavior of n – Window (regeneration time) statistics leads to the result that large amounts of randomly missing states are exponentially detrimental to the conveyed meaning [2,3]. Zipf – like tail rankings (dust) lead to interesting conclusions concerning frequency spectrum statistical techniques applied to language [1]. Reversibility for cellular types of algorithms in light of non – stationary techniques leads to similar conclusions [4]. It is the belief of this author that research into the relationship between ‘statistics’ (the curved portion of Figure 7) and ‘memory’ (the tail portion of Figure 7) holds fertile possibilities for advances in reversible (reconstruction) modeling, especially for very structured social phenomena such as language used in aviation communications.

5. REFERENCES

- [1] Han, Te Sun, *Information – Spectrum Methods in Information Theory*, Springer – Verlag, Berlin, 2003.
- [2] M.D. Moore and M.I. Savic, “Speech Reconstruction Using a Generalized HSMM (GHSMM),” *Digital Signal Processing Journal*, Elsevier Science, Vol. 14, No. 1, January 2004, pp. 37-53.
- [3] Moore, Michael D., *State Sequence Reconstruction Using a Generalized Hidden Semi Markov Model with Two Distinct Regeneration Times Applied to Speech*, Internal Publication (Doctor’s Thesis), Rensselaer Polytechnic Institute, Troy, NY, August 2004.
- [4] Wolfram, Stephen, *A New Kind of Science*, Wolfram Media, Champaign, IL, 2002.