Syllable Onset Detection from Acoustics

Michael Lee Shire

May 1997

Abstract

This paper describes a method of estimating the locations of syllable onsets in speech. While controversy exists on the precise definition of a syllable for American English, enough regularities exist in spoken discourse such that an operational definition will be correct a significant portion of the time. Exploiting these regularities, signal processing procedures extract indicative features from the acoustic waveform. A classifier uses these features to produce a measure of the syllable onset probability. Applying signal detection techniques to these probabilities yields segmentations that contain a large number of correct matches with true syllabic onsets while introducing an acceptable number of insertions. Higher level grammatical and linguistic knowledge is absent from the onset detection presented here. Reporting collaborative work with others in our research group, we show that the resulting segmentations can constrain and improve automatic speech recognition performance.

Contents

List of Figures

List of Tables

1 Introduction

The incorporation of syllabic and slow modulation information into speech recognition is a current research direction at the International Computer Science Institute (ICSI). Some researchers, such as Greenberg [9], have suggested the syllable as a basic unit of lexical access and stability in humans, particularly for informal, spontaneous speech. Some work has been done which considers modeling syllable-like units in lieu of phones for recognition [22, 18, 19, 21]. Various suprasegmental information such as prosodics is carried at the syllable level. Work by Wu and others continues to explore the use of syllable segmentation information to improve automatic speech recognition (ASR) [24, 14, 23]. Segmentation is a non-trivial source of information for pattern recognition tasks such as image scene analysis and speech recognition. Segmental information is one of the many potential information sources carried via the syllable.

Much of the previous research that estimates locations of syllables concentrates on detecting syllable nuclei, as in [17, 20]. The work described here attempts to directly estimate the location of the syllable onsets. Some ambiguity in the precise definition of a syllable provides an obstacle towards the use of a syllable in ASR, particularly for American English. The syllable structure of American English is considered by many to be complex. Rule-based definitions fail to account for all possible syllable realizations. Differences also exist between lexically canonical syllabification of words and the acoustic realizations of them, as noted in [16] for German. Greenberg defines a syllable as "a unitary articulatory speech gesture whose energy and fundamental frequency contour form a coherent entity" [8]. For practical reasons, syllables are typically described in terms of consonant-vowel structures such as CVC for "cat" and CCCVCCCC for "strengths", with a vowel or diphthong typically constituting the syllable nucleus. Though the structure of an American English syllable can be complex, recent statistical analysis of a spontaneous speech corpus reveals that the most frequently used words consisted of simple CV, CVC, VC, or V structures [10]. Similar observations were observed in telephone speech by Fletcher [7]. Whereas the syllable nuclei remain commonly

identifiable, the precise onsets become obscured in the presence of long strings of consonants. The common use of simple structures, however, provides regularities that may be exploited for syllable detection and segmentation. A set of methods developed for onset detection is described in this paper.

2 Overview

The onset detection technique reported here is adapted from the standard phoneme-based recognition system in use at ICSI. Signal processing schemes extract features from the acoustic speech signal. A Multi-Layer-Perceptron (MLP) uses these features as inputs for classification. The MLP is trained to distinguish between onset and non-onset frames. The system retains the same input features that we have used for phoneme classification. Additionally, a second set of acoustic features are used to provide additional indications of syllabic onsets. The MLP produces the probability that a given frame is a syllabic onset given the input acoustic features. A signal detection procedure uses these probabilities to determine the placements of the syllable onsets. This process operates directly on the acoustic waveform and incorporates no linguistic or grammatical knowledge (See Figure 1).

2.1 Test Corpus

A subset of the Numbers95 corpus [2] supplies a testbed for the experiments described here. The complete corpus comprises over 7000 continuous naturally spoken utterances excised from telephone conversations. The 92 words of the original corpus are numbers such as "twenty-seven" and "fifty".

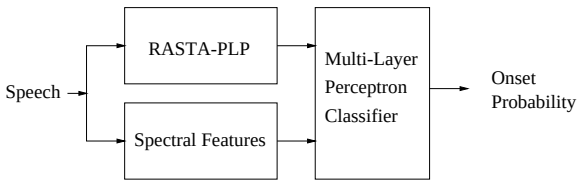

Figure 1: Overview of syllable onset detection.

The subset contains 33 words after eliminating ordinals such as "fifth" and most non-number words such as "dash." The selected subset utterances also have phonetic hand-transcribed labels; transcriptions are needed to provide a baseline for comparison. This subset is further divided into a training set, a cross-validation set, and a development set. The training subset contains 3590 utterances, the cross-validation subset contains 357 utterances, and the development subset contains 1206 utterances. The MLP training procedure uses frame classification performance on the cross-validation subset as an early stopping criterion. It computes an error score for the cross-validation set after each epoch of MLP weight training. Training stops when the error for this set reached its first local minimum. Early stopping prevents MLP parameters from over-fitting to the training data. The cross-validation set is also used for parameter tuning and for finding suitable thresholds for evaluation. The evaluation scores for the syllable onset detection use the cross-validation and development sets. The training, cross-validation, and development subsets contain utterances from different speakers.

Each utterance of the subsets has corresponding phonetic transcriptions hand-labeled by trained phonetic labelers at the Oregon Graduate Institute. The phone transcriptions are grouped into syllables using *tsylb2*, an automatic phoneme-based syllabification algorithm written by Bill Fisher at NIST [6]. An informal comparison with human syllabifications of spoken utterances suggest that *tsylb2* is a competent syllabifier [5]. For practical reasons, the *tsylb2* syllabification algorithm functions as the definition of a syllable. The frame corresponding to the start of the first phoneme of a given syllable phoneme group denotes a syllable onset. The hand-label derived segmentations serve as the ground truth for training and evaluation.

3 Feature Extraction

Syllable onsets are associated with synchronized rises in sub-band energy over adjacent sub-bands. Furthermore, the duration of the changes in energy is on the order of a syllable length. Figure 3

shows an example of the frequency-band energy envelopes for the utterance "seven seven oh four five" from the Numbers95 corpus. The envelopes are compressed to enhance where the envelope rises occur. We use features that emphasize these band energy changes for syllable onset detection. This contrasts with syllable segmentation algorithms such as those from Mermelstein [16] and Pfitzinger et. al. [17], which utilize local loudness and energy maxima and minima of the speech for demarking syllables.

The lengths of syllables vary with both stress and speaking rate, but are generally between 100 ms and 250 ms for an "average" speaker. This coincides with evidence that slow modulations in the range of 4 to 8 Hz are important for speech intelligibility [11, 4]. Figure 2 shows a histogram of syllable durations taken from the Numbers95 corpus subset. Here, the mode of the distribution is about 200 ms and the mean syllable duration is roughly 280 ms. The high mean is largely due to the nature and purpose of spoken numbers. Relatively important words tend to be spoken with more clarity and longer duration. Furthermore, the corpus has a restricted vocabulary with no short functional words which are commonly spoken very quickly.

Two sets of features derived solely from the sampled speech are used to detect syllable onsets. The MLP uses these features to produce the probability that a given frame corresponds to a syllable onset. Both features are described below.

3.1 Log-RASTA features

The first set of features used in detecting syllable onsets are the RASTA-PLP features [13]. Perceptual Linear Prediction (PLP) and RelAtive SpecTrAl (RASTA) analysis are front end feature extraction techniques used at ICSI and at other research facilities for standard phone-based speech recognition. PLP computes an auto-regressive spectral estimate of speech processed by an auditory model. RASTA performs bandpass filtering of the logarithm of the critical band trajectories. Figure 4 depicts the major processing steps for RASTA-PLP. First, spectral analysis separates the speech into critical-bands and the power spectrum is computed. The critical band values are

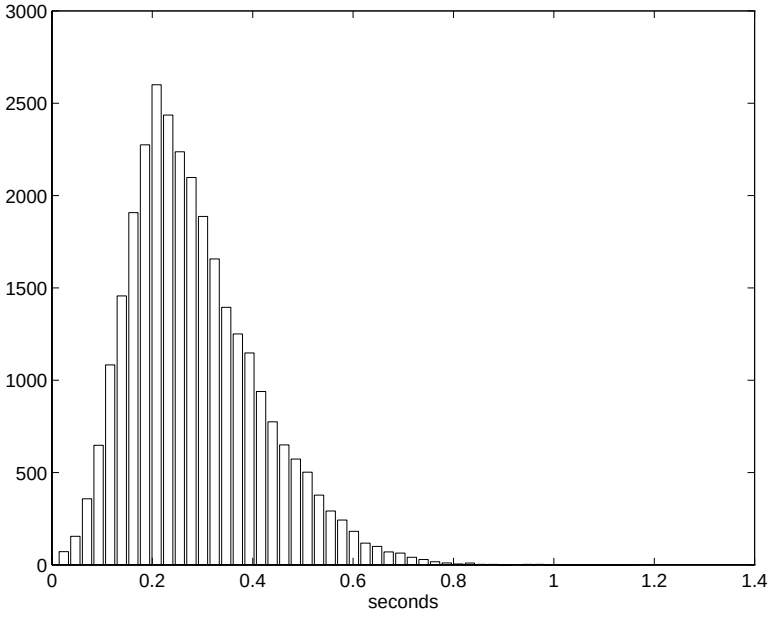

Figure 2: Histogram of Syllable Durations.

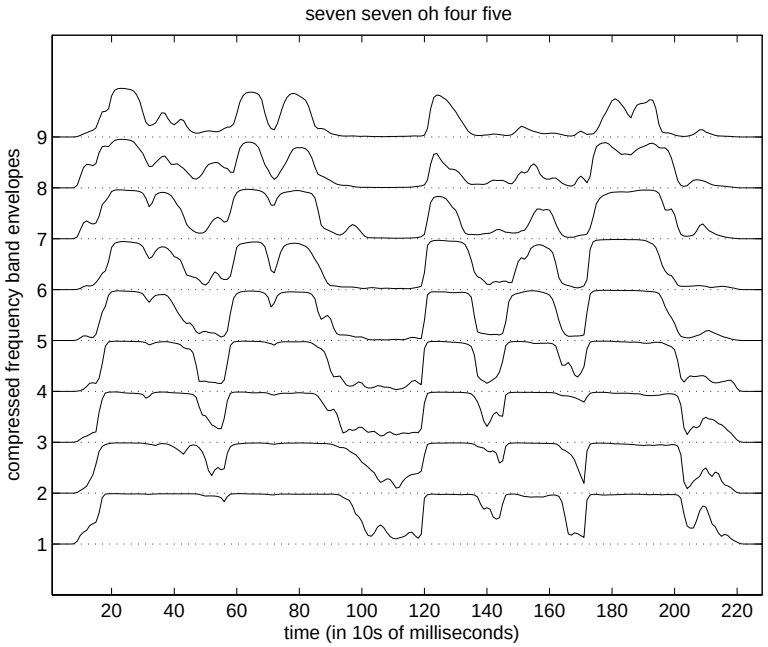

Figure 3: Compressed frequency band envelopes of the utterance "seven seven oh four five".

compressed with a logarithm and subsequently filtered with an IIR band-pass filter. The filtered values are then exponentiated and scaled with an approximation to the loudness curve and power law of human hearing. The resultant auditory power spectrum is modeled with an autoregressive (AR) model. Finally, cepstral coefficients are computed from the AR model.

Although primarily used for phone classification, RASTA-PLP incorporates desirable properties for syllable classification. The band-pass filter has the effect of emphasizing spectral change. In essence, it differentiates and re-integrates each band over time. Band-pass filtering helps capture the changes in band energy which we assume indicates boundaries of a syllable. Since the band-pass filter operates on the logarithm of the power, the filter also functions as a type of automatic gain control that can increase the relative strength of the energy in the consonants with respect to the typically stronger vowels. The emphasis helps reduce the effect of vowel onsets from dominating the response characteristics. The cepstral representation from a low-order AR model together with the critical band integration introduce a smoothing operation across the frequency axis. This helps capture the synchrony in neighboring frequency energies. For onset detection, we employ the energy and 8 RASTA-PLP cepstral coefficients with their derivatives as syllable onset features. The features are computed over a Hamming window of 25 ms of speech intervaled in 10 ms increments.

3.2 Spectral Features

Spectral onset features supplement RASTA-PLP. The spectral features attempt to locate gross regions of syllabic onsets by temporal processing in the power spectral domain. Our signal processing method, depicted in Figure 5, enhances and extracts the syllable onset properties described previously. The speech waveform is first decomposed into a spectrogram. Each time frame of the spectrogram is the squared magnitude of the 512 point Discrete Fourier Transform taken over a Hamming window of 25 ms of speech. The power spectrum is computed every 10 ms achieving the local frequency power spectrum versus time image. Fourth root compression and scaling yield the spectrogram image. An example of a spectrogram is shown in Figure 6.

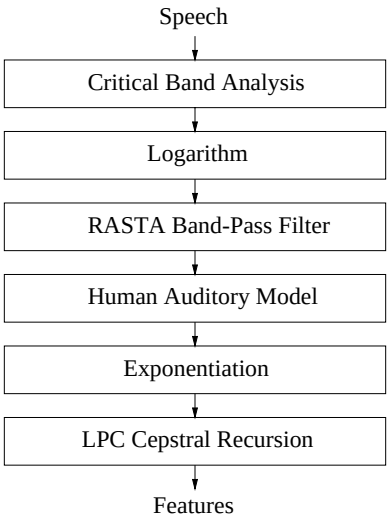

Figure 4: RASTA-PLP.

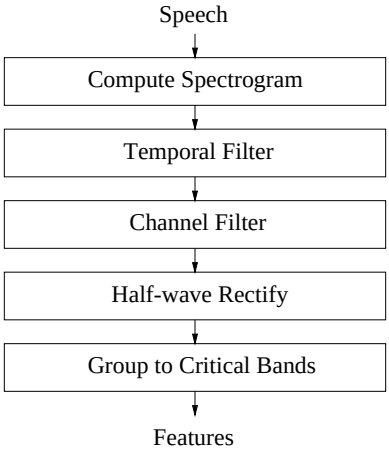

Figure 5: Major processing steps for the spectral onset features.

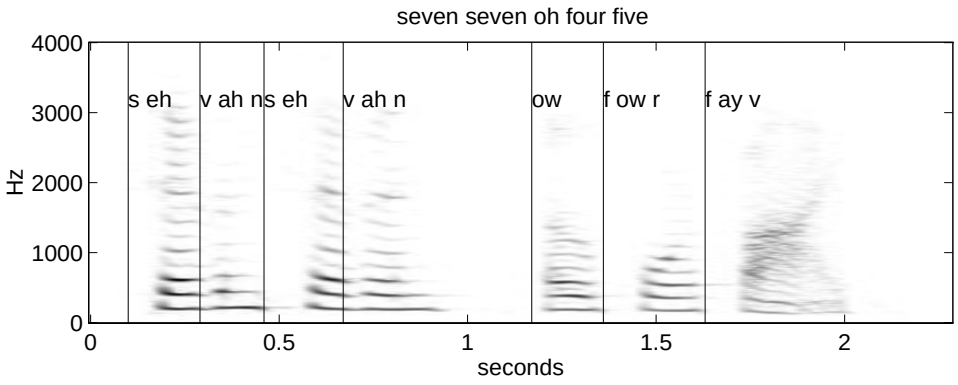

Figure 6: Spectrogram of the utterance "seven seven oh four five."

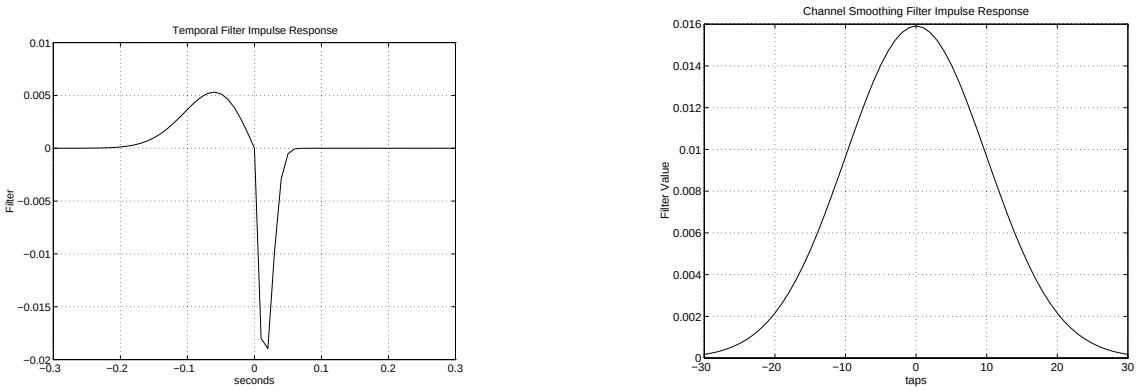

Figure 7: Temporal filter and channel filter.

The spectrogram is convolved with a temporal filter and a channel filter, effectively a two dimensional filter. The temporal filter, based on a Gaussian derivative, smoothes and differentiates along the temporal axis. The filter enhances changes in energy on the order of 150 ms, i.e. a short syllable length. The channel filter, a Gaussian, performs smoothing across the frequency channels, giving weight to regions of the spectrogram where adjacent channels are changing simultaneously. Figure 7 contains plots of the temporal and channel filters. The temporal and channel filters are similar to vertical edge detection filters in image processing. The filters have finite impulse responses and the channels are adjusted temporally to account for the average group delay.

Onsets are indicated by positive changes in energy. Half-wave rectification of the filtered spec-

trogram keeps only these positive changes. The frequency bands are subsequently averaged over nine critical band-like regions which have a frequency spacing derived from Greenwood's equation of the ear's frequency-position map [12]. The frequency band edges are shown in Table 1. The nine channels function as a set of syllable onset features. Figure 8 shows an example of the utterance "seven seven oh four five" after processing. Large values in the output correspond to possible syllabic onsets. The responses tend to peak in the regions prior to syllabic nuclei, which consist principally of vowels.

Edge	1	2	3	4	5	6	7	8	9	10
Freq (Hz)	203.1	312.5	437.5	609.4	812.5	1109.4	1484.4	1968.8	2625	3484.4

Table 1: Band Edges for Onset Feature Process.

The signal processing here bears many similarities to the RASTA-PLP processing. The differences reside principally in the temporal filtering and the threshold operation, which is absent in RASTA-PLP. The temporal filters for both processes have similar frequency responses but different temporal characteristics. Furthermore, they operate on different domains; RASTA-PLP operates in the log-power spectral domain. Finally, the spectral feature process lacks the human auditory-model scaling, such as equal loudness and power law, done in RASTA-PLP. In practice we have found that both sets of features complement one another for this application. In a pilot experiment, the combination of the two sets provided better results than either individually.

4 Syllable Onset Classification

A neural network classifier estimates the probability that a given frame of speech corresponds to a syllable onset. A three layer Multi-Layer-Perceptron (MLP) uses the features described above as input. A variant of the Error Back-Propagation Algorithm, commonly used at ICSI, trains the

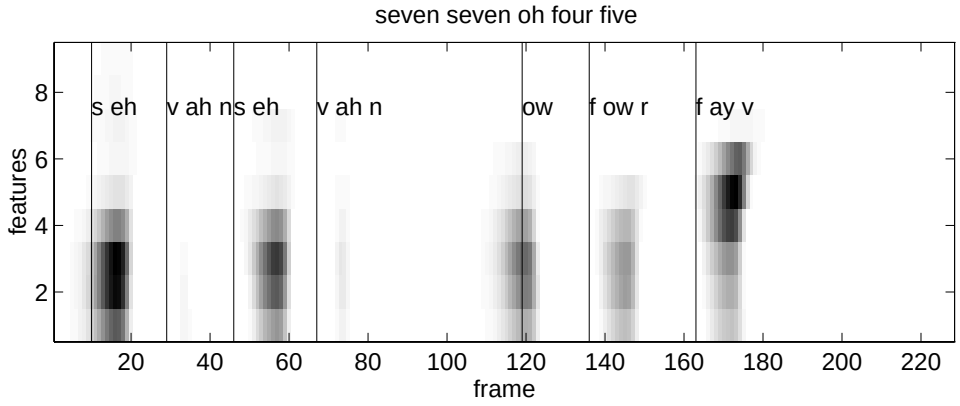

Figure 8: Example of utterance "seven seven oh four five" after processing.

weights of the MLP; weights are iteratively adjusted using a steepest descent procedure to minimize the relative entropy between the MLP output and the desired output. The input layer of the MLP consists of vectors of input features. To account for contextual effects, the input layer uses features from the current frame as well as the vectors of features for the four preceding and four following frames. With 27 features per frame, there are a total of 243 input nodes. The hidden layer contains 400 nodes. The output layer consists of 2 nodes corresponding to a syllable onset and a syllable non-onset.

The MLP is trained with a syllable onset tolerance window instead of the true syllable onset frame (Figure 9). The frame of the true onset and the ensuing four frames define a syllable onset window. The tolerance window broadens the single frame syllable onset to five frames. Effectively, the MLP is trained to recognize regions where a syllable onset can occur. Defining a region instead of a single frame helps correct possible variability of phonetic boundaries due to hand-labeling. Additionally, it increases the number of examples of the syllable onset target, thereby improving training and increasing output activation for the onsets.

The MLP training regimen uses features exclusively from the training subset of the NUMBERS95 corpus to adjust the MLP weights. Each epoch of training consists of iteratively updating the MLP connection weights for each training example. Training examples are presented in a random order

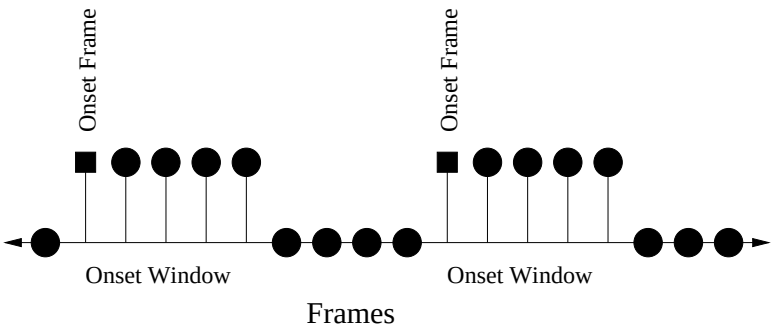

Figure 9: Syllable onset tolerance window.

to improve convergence. After each epoch, the training procedure computes the frame error rate for the cross-validation set. A frame error signifies that the MLP output is not closest to the correct target for the correct output. Target outputs are represented as a '1' or '0' depending on whether the corresponding frame is or is not in the syllable onset window. Training stops when the frame error rate reaches the first local minimum. Once trained, the MLP produces the probability estimate of a frame being an onset given the features for each frame of an input utterance. Figure 10 shows an example of the MLP output for the utterance "seven seven oh four five." The vertical lines denote the true syllable onsets for the utterance. This example shows peaks in the MLP probability output near where the true onsets occur. It also shows some extra peaks and high probability regions which do not correspond to true onsets. These other regions would count as false positive responses.

5 Evaluation

A signal detection procedure uses the MLP output probabilities to declare which frames are syllable onsets. For evaluation, frames are compared with the placement of the true syllable onsets derived from the hand-transcriptions. The procedure declares a match or hit if a true onset has at least one declared onset frame in its corresponding onset window. Again, the onset window consists of the frame containing the true onset and the four subsequent frames, as in Figure 9. The procedure counts a miss if there are no declared onset frames within four frames after a true onset. It counts

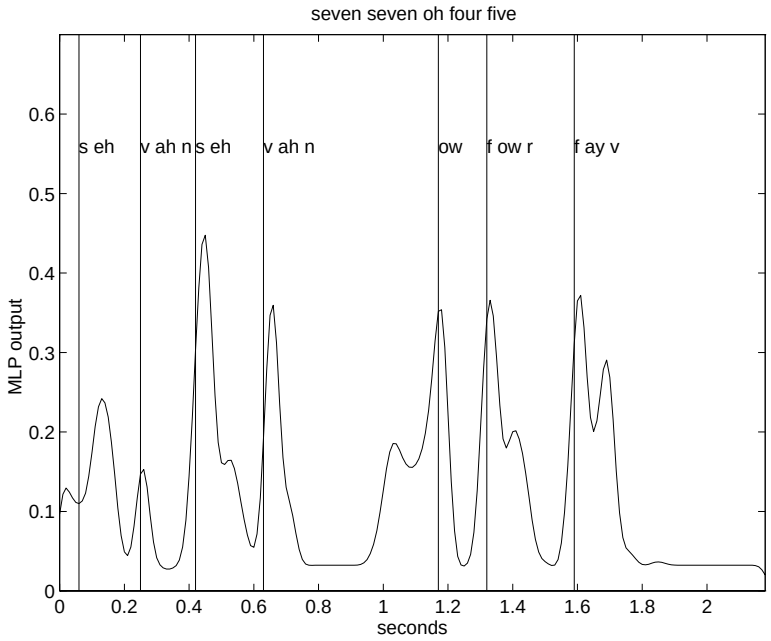

Figure 10: Example of MLP output.

an insertion for declared onset frames which do not match with a true onset.

5.1 Detection by Threshold

A simple approach to signal detection is to apply a threshold to the MLP outputs and perform a hit/miss analysis. Frames whose MLP outputs are above threshold are treated as onsets and those below are treated as non-onsets. Varying the threshold varies the number of true syllables which are matched and the number of false insertions which are introduced. The solid line in Figure 11 depicts a Receiver Operating Characteristic (ROC) curve for varying thresholds on the cross-validation set. Here the insertions are reported as an average number of false alarms per second. The number of hits and insertions are both inversely related to the threshold. An approach similar to a Neymann-Pearson formulation can determine a proper threshold for the development set. On the cross-validation set, the threshold is adjusted until a specified number of syllables are matched. This threshold is then applied to the development set to determine the number of hits,

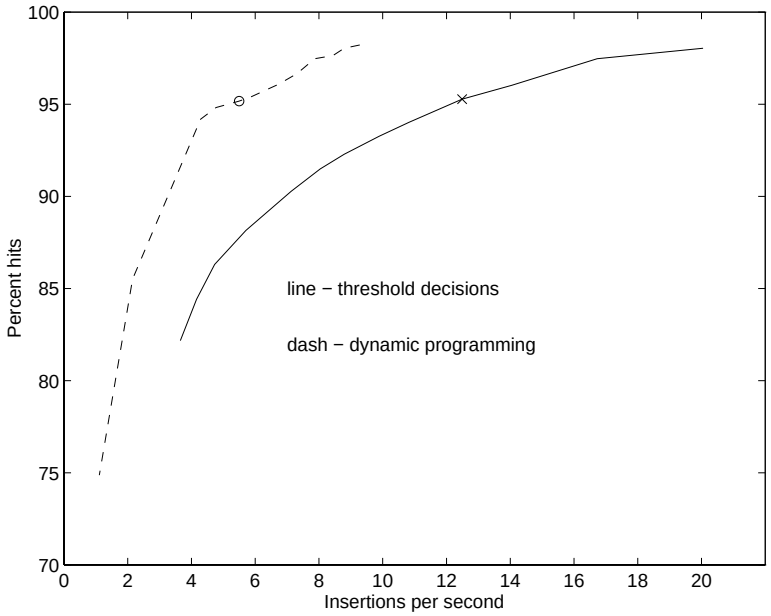

Figure 11: ROC curves.

misses and insertions.

Table 2 shows frame level scores for a threshold of 0.1291. This threshold is marked with an 'x' in Figure 11. Frame hits (misses) signify the number of declared onset (non-onset) frames which correspond to a syllable onset tolerance window frame. Insertions denote the number of declared onset frames which do not fall within a tolerance window. Non-onset matches correspond to the number of non-tolerance window frames which do not have declared onsets within them. The cross-validation set contains 1,739 syllables with 62,173 MLP output frames. The development set contains 5,975 syllables with 216,518 MLP output frames. Table 3 shows the percentage of syllables that have at least one declared onset within their respective tolerance windows.

5.2 Adding Minimum Duration Constraint

The MLP output varies smoothly over time. The threshold criterion therefore typically declares clusters of frames as onsets. This causes the average number of false alarms per second to increase. Requiring a minimum number of non-onset frames between any two onsets can reduce the number

Subset		**Frame Hits**	**Frame Misses**	**Insertions**	**Non-onset Matches**
Cross Validation	Frames	7313	1332	7768	45760
Thresh = 0.1291	Percent	84.59%	15.41%	**14.51%**	85.49%
Development	Frames	24986	4884	26373	160275
Thresh = 0.1291	Percent	83.65%	16.35%	**14.13%**	85.87%

Table 2: Frame-level Hits, Misses, and Insertions.

Subset	**Percent Hits**	**PercentFrame Insertions**
Cross Validation	95.28%	14.51%
Development	94.21%	14.13%

Table 3: Syllable onset hits and frame insertions.

of false insertion frames. For example, among the true onsets, there are no examples of two distinct onsets being in adjacent frames. Disallowing a multiple number of detected onsets from occupying the same minimum duration window reduces the number of detected onset frames, and hence the number of frame insertions.

One method of imposing the minimum duration constraint is with a Viterbi search using a Hidden Markov Model formulation [3, 1]. Here, dynamic programming finds the path which *best* matches or produces the least *cost* path with a syllable model, such as the one in Figure 12. The syllable model consists of a sequence of states with permissible transitions and transition costs between the states. States correspond to syllable onsets or non-onsets. For each utterance, a lattice is generated with the abscissa corresponding to the frames of the utterance and the ordinate corresponding to the states of the syllable model. Each frame/state pair in the lattice has associated with it a local cost and a transition cost. The local cost consists of the negative logarithm probability of the MLP output for that state. The two MLP outputs are divided by their respective training prior probabilities before computing the cost. The syllable model constrains the allowed frame/state transitions and specifies the cost associated with making each transition. The transition costs consist of the negative logarithm of the transition probabilities. The *Viterbi* algorithm finds the least cost path for each utterance. Figure 13 illustrates a sample where the least cost path depicts two syllable onsets.

To satisfy minimum duration constraints, a chosen syllable model requires syllable onsets to be separated by a minimum number of frames. For the syllable model depicted in Figure 12, the minimum separation between syllables is five frames (50 ms). The out-going transition probabilities were arbitrarily chosen to be 0.5 for all states except for the state corresponding to the onset and the final right-most state. Table 4 shows evaluation scores for the model depicted in Figure 12. Modification of the transition probabilities changes the sensitivity and frequency with which syllable onsets are declared. The dashed line in Figure 11 shows the ROC curve for the cross-validation set using the previous syllable model. Varying the transition probabilities for the right-most states traces the curve.

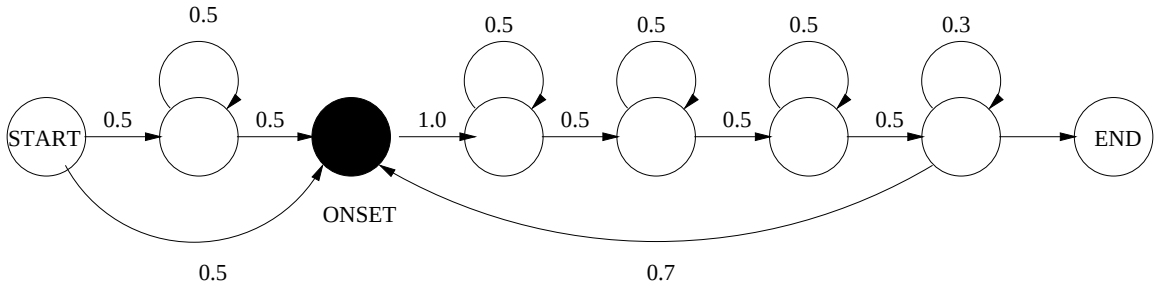

Figure 12: Syllable model for dynamic programming.

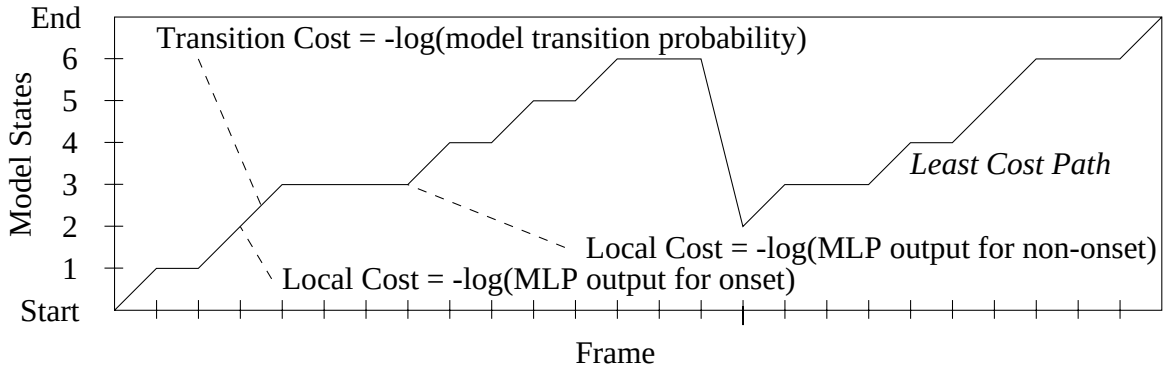

Figure 13: Illustration of least cost Viterbi path.

Subset	Percent Hits	Percent Frame Insertions
Cross Validation	95.17%	6.38%
Development	94.53%	6.28%

Table 4: Dynamic programming duration constraint results.

System	Error Rate sub./ins./del.
Single pronunciation lexicon no onset information	10.8% 5.8%/3.1%/1.8%
Single pronunciation lexicon Viterbi onset information	7.3% 4.9%/0.9%/1.5%

Table 5: Comparison of systems using a single pronunciation lexicon with and without *cheating* onset boundaries.

5.3 Application to Speech Decoding

The syllable onset information obtained via the threshold criterion was incorporated into a syllable-based speech decoder developed by Wu [24]. This decoder incorporated a syllable lexicon and used the onset information to constrain the regions where a syllable may begin. The decoder showed statistically significant improvement in word recognition over a baseline decoder which did not use *cheating* syllable onset information. The *cheating* boundaries were obtained from forced Viterbi alignment with the word transcriptions. Using a single-pronunciation lexicon, word error reduced 38% relative to a baseline decoder which did not use *cheating* syllable onset information (Table 5). This represents an upper bound indication of how much onset information can improve recognition performance. Incorporation of the acoustically derived syllable onsets from a threshold detection criterion resulted in some improvement in recognition results. Using a multiple-pronunciation lexicon, errors were reduced by 10% relative to the baseline (Table 6).

The baseline decoders were allowed to hypothesize a syllable onset at every frame of an utterance. The syllable-based decoder was allowed to hypothesize syllable onsets if the separate onset detection scheme declared an onset within the tolerance window of 5 frames. Using the threshold criterion with

System	Error Rate sub./ins./del.
Multiple pronunciation lexicon no onset information	9.1% 5.3%1.3%2.4%
Multiple pronunciation lexicon acoustic onsets information	8.2% 4.8%1.3%2.1%

Table 6: Comparison of systems using multiple pronunciation lexicon with and without acoustic onsets from syllable detection.

a threshold of 0.12 on the MLP output, 58% of the frames in the development set were eliminated from consideration as a potential syllable onset.

The experiment with the *cheating* boundaries and the experiment with the acoustic boundaries used different syllable lexical for practical reasons. The principal reason was that the acoustic onsets did not align adequately with the canonical single-pronunciation lexicon. Additionally, applying the multiple-pronunciation lexicon to the Viterbi procedure to produce aligned onsets would have significantly increased the complexity of decoding. The results are therefore not directly comparable. Both experiments do, however, provide an indication of the improvement from adding syllable onset constraints to decoding.

6 Conclusion

The syllable onset detection method presented here seeks to estimate the locations of syllabic onsets from acoustic information alone. It does not directly incorporate lexical or grammatical knowledge. The basic premise of the method is to exploit the relationship between syllable onsets and rises in energy in adjacent frequency channels. Using various signal detection criterion, the analysis

demonstrates that an acoustic criterion alone can achieve a strong number of hits with an acceptable number of insertions. Furthermore, insertions can be decreased by the addition of duration constraints. While maintaining roughly 94% hits, a dynamic programming method for constraining inter-syllable occurance reduces the number of insertions by as much as 60%. Additional measures such as region matching with syllable nuclei or other onset detection techniques are likely to improve performance.

The major impetus for locating syllable onsets is to add constraints to a speech recognition system by limiting where syllable onsets can be hypothesized. Experiments with such onset constraints demonstrate improvement in speech decoding. Further, even simple threshold criterion detection can eliminate roughly 60% of speech frames from consideration as an onset. This is with a widening tolerance window of 50ms and without benefit of duration constraints. Incorporation of the acoustic-derived segmentation into the decoding process has illuminated some discrepancy between concepts of acoustic-phonetic and phonological representations of syllables. This discrepancy is often apparent in word sequences where the coda of the first word is consonantal and the onset of the following word is vocalic. For example, the word sequence "five eight" has a phonological or canonical representation of /fayv/ /eyt/ while the phonetic realization is more typically [fay][veyt]; here the /v/ in the first syllable appears as part of the second.[1] Difficulties also arise from ambisyllabicity where the precise boundary between two adjacent syllables is ambiguous. This occurs frequently where the same phone appears in both the coda of the first syllable and the onset of the second, as in "four eight" ([foh[r]eyt]) and "nine nine" ([nay[n]ayn]). Consequently, the boundary of the syllables is within the phone and difficult to locate with precision. The ground-truth segmentations do not explicitly reconcile ambisyllabicity in the experiments here. This might have introduced shortcomings in the MLP training. Regardless, the onset detection technique shows promise as an additional information stream to a speech recognition system.

[1]Many such coarticulation effects are chronicled within the *Sandhi* framework by Panini. A treatment can be found in [15].

7 Acknowledgments

I would like to thank and acknowledge the following people. Nelson Morgan guided the overall progression of this work. Steven Greenberg provided expert testimony on properties and analyses of syllables. Su-Lin Wu created the syllable-based speech decoder which provided the focal application for this work. Dan Ellis provided an adaption of the *tsylb* algorithm used for the baseline syllable boundaries. Nikki Mirghafori gave me valuable comments and discussions. Jeff Bilmes, Dan Gildea, Eric Fosler, and Brian Kingsbury provided some technical assistance and discussions. I would also like to thank Lokendra Shastri for reviewing this paper and the people at ICSI for providing a great research atmosphere. This work was funded through a Department of Defense subcontract from the Oregon Graduate Institute.

References

[1] H. Bourlard and N. Morgan. *Connectionist Speech Recognition- A Hybrid Approach*. Kluwer Academic Press, 1994.

[2] Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. Numbers corpus, release 1.0, 1995.

[3] J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*, chapter 10–14. Macmillan Publishing Company, New York, 1993.

[4] R. Drullman, J. M. Festen, and R. Plomp. Effect of temporal envelope smearing on speech reception. *JASA*, 94(2):1053–1064, Feb. 1994.

[5] D. Ellis, 1996. Personal communication.

[6] B. Fisher. The *tsylb2* program, Aug. 1996. National Institute of Standards and Technology Speech.

[7] H. Fletcher. *Speech and Hearing in Communication*. Krieger, 1953.

[8] S. Greenberg, 1996. Personal communication.

[9] S. Greenberg. Understanding speech understanding: Towards a unified theory of speech perception. In *Proceedings of the ESCA Workshop (ETRW) on The Auditory Basis of Speech Perception*, pages 1–8, Keele, United Kingdom, July 1996. ESCA.

[10] S. Greenberg. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop (ETRW) on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32, Pont-á-Mousson, France, Apr. 1997. ESCA.

[11] S. Greenberg and B. E. D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. In *ICASSP*, volume 3, pages 1647–1650, Munich, Germany, April 1997. IEEE.

[12] D. D. Greenwood. Critical bandwidth and the frequency coordinates of the basilar membrane. *JASA*, 33:1344–1356, 1961.

[13] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, Oct. 1994.

[14] M. Jones and P. Woodland. Modelling syllable characteristics to improve a large vocabulary continuous speech recogniser. In *ICSLP*, volume 4, pages 519–522, Yokohama, Japan, Sept. 1994.

[15] E. M. Kaisse. *Connected Speech: The Interaction of Syntax and Phonology.* Academic Press, 1995.

[16] P. Mermelstein. Automatic segmentation of speech into syllabic units. *JASA*, 58(4):880–883, Oct. 1975.

[17] H. R. Pfitzinger, S. Burger, and S. Heid. Syllable detection in read and spontaneous speech. In *ICSLP*, volume 2, pages 1261–1264, Philadelphia, Pennsylvania, Oct. 1996.

[18] B. Plannerer and B. Ruske. Recognition of demisyllable based units using semicontinuous Hidden Markov Models. In *ICASSP*, pages I581–I584, San Francisco, California, Mar. 1992.

[19] B. Plannerer and B. Ruske. A continuous speech recognition system using phonotactic constraints. In *Eurospeech*, pages 859–862, Berlin, Germany, Sept. 1993.

[20] W. Reichl and G. Ruske. Syllable segmentation of continuous speech with artificial neural networks. In *Eurospeech*, pages 1771–1774, Berlin, Germany, Sept. 1993.

[21] G. Ruske, B. Plannerer, and T. Schultz. Stochastic modeling of syllable-based units for continuous speech recognition. In *ICSLP*, pages 1503–1506, Banff, Canada, Oct. 1992.

[22] K. Shinoda and T. Watanabe. Unsupervised speaker adaption for speech recognition using demi-syllable HMM. In *ICSLP*, volume 2, pages 435–438, Yokohama, Japan, Sept. 1994.

[23] Y. Wakita and E. Tsuboka. State duration constraint using syllable duration for speeech recognition. In *ICSLP*, volume 1, pages 195–198, Yokohama, Japan, Sept. 1994.

[24] S.-L. Wu, M. L. Shire, S. Greenberg, and N. Morgan. Integrating syllable boundary information into speech recognition. In *ICASSP*, volume 2, pages 987–990, Munich, Germany, April 1997. IEEE.