Speech Recognition Using Syllable-Like Units

Zhihong Hu, Johan Schalkwyk, Etienne Barnard, Ronald Cole
(zhihong,johans,barnard,cole)@cse.ogi.edu
Center for Spoken Language Understanding,
Oregon Graduate Institute of Science and Technology

ABSTRACT

It is well known that speech is dynamic and that frame-based systems lack the ability to realistically model the dynamics of speech. Segment-based systems offer the potential to integrate the dynamics of speech, at least within the phoneme boundaries, although it is difficult to obtain accurate phonemic segmentation in fluent speech. In this paper we propose a new approach which uses syllable-like units in recognition. In the proposed approach, syllable-like units are defined by rules and used as the basic units of recognition. The motivation for using syllable-like units is (1) by modeling perceptually more meaningful units, better modeling of speech can be achieved; and (2) this method provides a better framework for incorporating dynamic modeling techniques into the recognition system. The proposed approach has achieved the same recognition performance on the task of recognizing months of the year as compared to the best frame-based recognizer available.

1. Introduction

Many current speech-recognition systems use phonemes or phoneme-like units to model the basic sounds of speech. The advantages of using such subword units are (1) phonemes are linguistically well defined; therefore pronunciation models based on phonemes for virtually any word can be looked up easily from a dictionary; (2) pronunciation variability due to linguistic context, accent or dialogues can be easily represented by applying rules to base forms; (3) the number of units is small; for example, there are about 50 phoneme-like units for English. Phoneme-like units therefore typically require significantly less data to train than would be needed for whole word modeling (for example).

However, it is well known that speech is non stationary. In fluent speech phoneme segmentation is not easy as is shown (for example) in the process of labeling fluent speech, some of the segment boundaries are arbitrary and must be defined by convention [1].

Phoneme-based recognition has been attempted using both frame-based and segment-based approaches. Frame-based systems are currently more popular since they do not require explicit detection of segment boundaries and thus give better classification performance. However, they suffer from severe modeling limitations. Speech is modeled as a sequence of conditionally independent frames, and these are assumed to have piecewise-constant statistics. Frame-based systems thus lack the ability to model the dynamics of speech realistically. Although segment-based systems have the ability to integrate the dynamics of speech, this is mostly constrained within the phoneme boundaries. Given the difficulty of obtaining accurate phonemic segmentation, especially in fluent speech, this is particularly restrictive.

Various attempts have been tried to overcome the limitations of frame-based systems [2, 3]. However, these attempts are still constrained by the phoneme-based paradigm, therefore leaving the fundamental problems unsolved.

Recent research in segment based systems [4, 5] shows how to model both the dynamics within the segment boundary as well as the transitional part between segment boundaries. Various trajectory models are used, which proved to be very useful at modeling dynamics in speech. However, these approaches either need accurate segmentation, which is difficult to achieve, or assume every single frame as a possible boundary, followed by an expensive exhaustive search method applied to a lattice constructed using all the possible boundaries [4].

In order to take advantage of the modeling ability of the trajectory models and not be constrained by the segmentation accuracy, we propose a new recognition strategy which uses syllable-like units as the basic unit for recognition. By modeling a syllable-like unit we can capture not only the dynamics within phoneme boundaries but also the dynamics in the transition region. Many of the transitions between phonemes (vowel-vowel or liquid-vowel transitions) are difficult to detect using current segmentation algorithms. Grouping these phonemes together, forming a new unit to be modeled, the effect of boundary deletions will therefore be less of a problem in this proposed recognition system.

In this paper, Section 2 describes the proposed algorithm in

more detail. Section 3 presents the experimental setup and results obtained. Finally Section 4 concludes and presents some future work towards improving this new recognition paradigm.

2. Recognition Paradigm

Figure 1 depicts the outline of the proposed recognition paradigm.

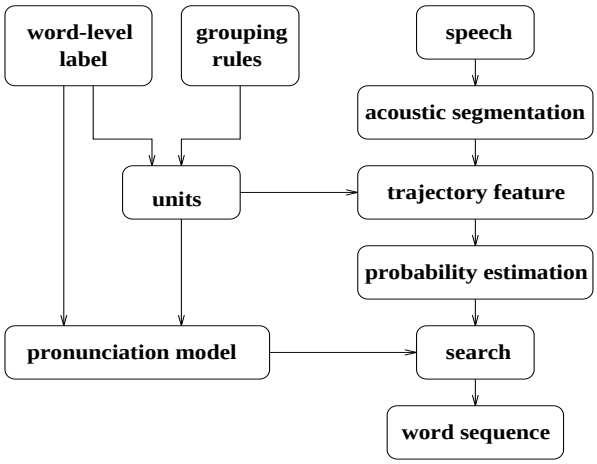

Figure 1: System Paradigm.

We define the criterion of grouping phonemes into new syllable-like units as follows: Phoneme sequences for which the boundary is difficult to detect are grouped together forming a new set of base recognition units. In the experiments performed we study the merging of a stop and the following vowel, or a vowel and the adjacent vowel or semivowel. After syllable-like units are defined according to the set of predefined rules, word pronunciation models are generated using these units.

A simple and computationally efficient segmentation algorithm is used to produce segmentation for the incoming speech. The parameters of the algorithm are tuned to generate a segmentation containing all boundaries where the spectrum changes rapidly. Comparing the machine segments with hand-labeled segments, we observe both insertion of segments, as well as deletion of phonemic boundaries. The deletion of phonemic boundaries occurs mostly within the regions which are in general difficult to detect (e.g. vowel-vowel boundaries). Because the syllable-like units are defined to ignore these boundaries, phonemic-boundary deletion should not affect the recognition performance. Also, since the segmentation algorithm in general over-generates segments, almost all the boundaries between units defined in this approach will be detected. Using this segmentation algorithm, insertions of boundaries will therefore occur.

Statistical trajectory models as described in [4] are computed for each of the units defined. Artificial neural networks or Gaussian mixture models are then trained to estimate probabilities for the units defined.

The search is implemented using the Viterbi algorithm in a time-asynchronous manner. This Viterbi algorithm therefore only considers the boundaries detected by the segmentation algorithm as hypothesized unit boundaries. Having the knowledge of how many insertions the segmentation algorithm may generate, only a limited number of segments of look-ahead are considered for each state in the search algorithm, which produces a significant saving in time. Figure 2 depicts the asynchronous nature of the Viterbi algorithm implemented.

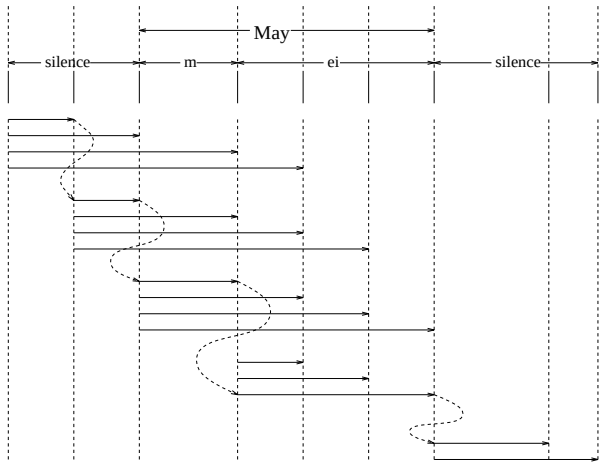

Figure 2: Time-asynchronous Viterbi search, using a segment look-ahead of three. The silence model is constrained to a one segment look-ahead. The connected path indicates the highest scoring path for the word May, pronounced (m ei).

3. Experiments and results

We tested our approach using a database consisting of the twelve months of the year. This database is a subset of the Census data collected at OGI [6]. [1]

The database consists of a total of 814 transcribed files for the training set, and 796 transcribed files for the development test set. This results in an average of 67 repetitions per word in the vocabulary. Due to the multi-syllabic nature of the words in the database, this task is well-suited for testing the dynamic modeling ability of the syllable-like units, as well as the algorithms needed to choose the basic recognition units.

For each unit, the trajectory model consists of 10 states. [2]

[1] For more information on the census database contact noel@cse.ogi.edu.

[2] In Goldenthal's thesis [4], he found that 10 states is good enough to model a segment. Our experiments also show that

The feature space consists of 8 PLP coefficients for each state plus log duration of the unit. Therefore the total dimension of the feature vector is 81. For each of the experiments presented we train neural networks to be probability estimators for the base recognition units.

Multiple Gaussian models were also trained as probability estimators, but they have not yet been successfully integrated into the recognition system. The reason for this is that insufficient data are available to train a Gaussian model even when using diagonal covariance matrices (for some of the categories, only about 70 examples for each unit are available.). The need for more data to train good Gaussian models may limit the feasibility of this approach.

In the following sections we discuss the series of the experiments performed.

3.1. Segment-based system using trajectory models

In this experiment, phoneme is used as the basic unit in recognition. Table 1 depicts the pronunciation models for the twelve words.

january	dZ @ n [j] u 3r i:
february	f E [vc] b [9r] [j] u 3r i:
march	m A 9r uc tS
april	ei [uc] ph 9r I l
may	m ei
june	dZ u n
july	dZ u l aI
august	A vc g ^ s uc th
september	s E uc [ph] [uc] th E m vc b 3r
october	A uc [kh] [uc] th oU vc b 3r
november	n oU v E m [vc] b 3r
december	d i: s E m [vc] b 3r

Table 1: Pronunciation models for the twelve months of the year. Phonemes are represented in the worldbet symbol set.

Since only the word-level transcriptions are available, trajectory models for phonemes are trained and tested based on the forced aligned phoneme boundaries, generated using a general purpose frame-based recognizer. Given the correct phonemic segmentation we would expect the trajectory models to perform relatively well on this task. Table 2 depicts the result obtained.

	Phoneme level	Word level	
Segmentation	Aligned	Aligned	Automatic
Error Rate	16.7%	0.08%	73.2%

Table 2: Experimental results using phoneme as the basic unit for segment-based recognition.

more states do not help to improve the performance.

Given the forced aligned boundaries we can achieve a phoneme-level (30 phonemes) recognition accuracy of 83.3%. The word recognition rate given the forced aligned boundary is 99.2%. However, when using the boundaries generated automatically, the word recognition accuracy dropped to 26.8%. The conclusion that can be drawn from this result is that given an accurate segmentation, the trajectory model can capture the information necessary to perform good recognition. However, inaccurate segmentation will significantly degrade performance.

3.2. Using syllable-like units in the recognition

In this experiment, we group some of the phonemes to form a new set of recognition units. The grouping is performed according to the rules described in section 2. Table 3 depicts the chosen base recognition units, with the resulting pronunciation models shown in table 4.

ja	dZ @	nua	n [j] u	ry	3r i:
fe	f E	vc	vc	brua	b [9r] [j] u
m	m	ar	A 9r	uc	uc
ch	tS	a	ei<ph	pril	ph 9r I l
ay	m<ei	ju	dZ u	ne	n
ly	l aI	au	A>g	gu	vc g ^
s	s	t	th	se	s E
te	th E	m2	m [vc]	ber	b 3r
o	A>k	k	uc [kh]	to	th oU [vc]
no	n oU	ve	v E	de	d i:

Table 3: Syllable-like units chosen according to the set of predefined rules.

january	ja nua ry
february	fe [vc] brua ry
march	m ar [uc] ch
april	a [uc] pril
may	m ay
june	ju ne
july	ju ly
august	au gu s [uc] t
september	se uc te m2 ber
october	o k [uc] to [vc] ber
november	no ve m2 ber
december	de se m2 ber

Table 4: Pronunciation models for the twelve months in a year, using the syllable-like units defined.

Trajectory models are trained based on the forced aligned boundaries. As in the first experiment performed we test the system based on the given forced aligned segmentation as well as the automatically generated segmentation. The results are shown in table 5.

The classification rate of the syllable-like units (29 units)

	Unit level	Word level	
Segmentation	Aligned	Aligned	Automatic
Error Rate	15.2%	3.5%	5.2%

Table 5: Experimental results using syllable-like units for segment-based recognition.

is 84.8%. The word recognition rate based on the forced aligned boundaries is 96.5%, which is somewhat lower than the phoneme based system. However, word recognition rate based on the automatic segmentation is 94.8%. This result shows that by choosing bigger than phoneme units this proposed approach is much less sensitive to segmentation accuracy as compared to a phoneme-level segment-based recognition system.

During this experiment, we also found that adding negative training data for each unit(class) in the training process can help to improve the probability estimator's rejection capability when the input is not one of the units in the chosen unit set. This produced a 46.4% error reduction in performance (word error dropped from 9.7% to 5.2%).

3.3. Iterative refinement

Trajectory models above are computed from unit boundaries obtained using the forced-aligned boundaries of a general purpose frame-based recognizer. However, the frame-based recognizer alignment does not correspond well to the true acoustic unit boundaries. In this experiment, we use the syllable-like recognizer to generate the unit boundaries, from which a new set of trajectory models are computed. Table 6 depicts the recognition results obtained.

	Unit level	Word level
Segmentation	Re-Aligned	Automatic
Error Rate	12.5%	4.5%

Table 6: Experimental results after training the trajectory models based on the realigned unit boundaries obtained using the syllable-like recognizer of the experiment 2.

These results indicate that unit boundaries computed from the realignment correspond more accurately to the true unit boundaries than what we found using the general frame-based recognizer. Having more accurate unit boundaries in turn produced better models resulting in higher recognition accuracy. This result is comparable to our current best frame-based recognizer (with an error rate of 4.0%).

4. Summary and Future Work

Using syllable-like units, we are able to design units which encapsulate the dynamics of speech. Furthermore, since these units typically contain the phoneme boundaries which are in general difficult to detect, boundary deletion due to automatic acoustic segmentation has less severe effects on the performance than it would have on phoneme-based segmental recognition systems. The proposed algorithm also provides a better environment for investigating models which describe the dynamics of speech.

Each component of the proposed recognition paradigm has potential for improvement. The following are some of these areas:

- Extensions of this new method needs to be investigated for a continuous recognition task (such as digits) which has enough data to train the trajectory models. With more data we will be able to compare Gaussian mixture models and neural networks. Currently due to the lack of training data only neural networks are able to learn the underlying unit patterns.
- Currently all trajectory models consist of a fixed number of states. Due to the varying length of the chosen units, this might lead to either over- or under-sampling of the underlying trajectories. Therefore we need to investigate the possibility of having a variable number of states for units having different length.
- In the current implementation we do not have the ability to perform out-of-vocabulary rejection. A good garbage model would alleviate this deficiency. For example, if Gaussian models can be trained, it could possibly yield better garbage rejection ability than neural networks.

Acknowledgement: This work was supported through a grant in the Young Investigator Program of the Office of Naval Research. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

5. REFERENCES

1. R. Cole, B. Oshika, M. Noel, T. Lander, and M. Fanty, "Labeler agreement in phonetic labeling of continuous speech," in *International Conference on Spoken Language Processing*, pp. 2131–2134, Sept. 1994.
2. O. Ghitza and M. Sondhi, "Hidden markov models with templates as non-stationary states: an application to speech recognition," *Computer Speech and Language*, vol. 2, pp. 101–119, 1993.
3. Z. Hu, E. Barnard, and R. Cole, "Transition-based feature extraction within frame-based recognition," in *Eurospeech*, pp. 1555–1558, October 1995.
4. W. Goldenthal, *Statistical Trajectory Models for Phonetic Recognition*. PhD thesis, M.I.T., Auguest 1994.
5. M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Transaction on Accoustics, Speech, and Signal Processing.*, vol. 37, no. 12, pp. 1857–1869, 1989.
6. R.A.Cole, D.G.Novick, M.Fanty, P.Vermeulen, S.Suttong, D.Burnett, and J.Schalkwyk, "A prototype voice questionary for the US census," in *International Conference on Spoken Language Processing*, Sept. 1994.