ADVANCES IN ALPHADIGIT RECOGNITION USING SYLLABLES

Jonathan Hamaker, Aravind Ganapathiraju, Joseph Picone, John J. Godfrey

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University, Mississippi State, Mississippi 39762
{hamaker, ganapath, picone}@isip.msstate.edu, godfrey@csc.ti.com

ABSTRACT

***Abstract* - In this paper, we present a set of experiments which explore the use of syllables for recognition of continuous alphadigit utterances. In this system, syllables are used as the primary unit of recognition. This work was motivated by our need to verify and isolate phenomena seen when performing syllable-based experiments on the SWITCHBOARD corpus. The performance of our base syllable system is better than a crossword triphone system while requiring a small portion of the resources necessary for triphone systems. All experiments were performed on the OGI Alphadigits corpus, which consists of telephone-bandwidth alphadigit strings. The WER of the best syllable system (context-independent syllables) reported here is 11.1% compared to 12.2% for a crossword triphone system.**

1. BACKGROUND

A robust and reliable alphadigit system has long been a goal for automated telephone transactions. Recent work on both alphabet and alphadigit systems has focused on resolving the high confusion rates for the E-set (B, C, D, E, G, P, T, V, Z, THREE), A-set (A, J, K, H, EIGHT). Further, for telephone bandwidth data, the S-F confusion pair is also important.These problems occur mainly because the acoustic differences between the letters of the sets are minimal. Even humans have trouble making such distinctions in the telephone environment due to bandwidth and microphone constraints.

1.1. State-of-the-Art in Alphadigit Recognition

Phoneme-based models, when trained in a strongly supervised mode, are capable of capturing phonetic detail. This is especially true in the case of context-dependent phonemes. In the case of spoken letters, discriminating information exists for a short duration in the form of glottal stops, typically at the onset of the word. Most state-of-the-art systems therefore incorporate modeling of onset segments, spectral transitions, and the use of letter-dependent models [1]. Detailed phoneme modeling however introduces a number of challenges such as a tremendous increase in complexity and search space, not to mention the requirement to have narrowly transcribed speech data. Since the alphadigit problem is essentially a small vocabulary recognition task, the growth in complexity of such systems can be mitigated by using longer acoustic models. A syllable-sized unit is one such unit — extremely stable and well-suited for simultaneous temporal and spectral modeling. With syllables, these fine-grain modeling approaches may become unnecessary since a longer context will allow the models to learn these events statistically. Recent work on using syllables for large vocabulary tasks [2] has shown promise.

Most research in the last twenty years on continuous telephone alphadigit recognition has been centered around phone-based, speaker-dependent systems. Word error rates (WER) for such systems [3,4] are typically in the 10% range. Speaker-independent technology [5] lags speaker-dependent technology with published error rates extending to 20%. In general, alphabet recognition is a much more difficult task than digit recognition. State-of-the-art connected telephone digit recognition performance is typically less than 1% WER.

1.2. OGI Alphadigit Corpus

The OGI Alphadigit corpus [6] is a recent release and has many things in common with the SWITCHBOARD corpus [7] (SWB). It is a telephone database collected using a T1 interface to the telephone network. There are over 3000 subjects in the corpus. Each was given a list of either 19 or 29

alphanumeric strings to speak. The strings in the lists were each six words long, and each list was designed to balance the phonetic context of all letter and digit pairs. There were 1102 unique prompting strings.

Since there have been no published results on this data, there exists no standard partitioning of the database for common evaluations. Hence, we have developed such a partitioning by splitting the data along gender lines. Table 1 shows the separation of the training and testing data. In addition we have defined a 3000 utterance evaluation set from the test data, on which all of our results are quoted. This test set definition [11] is publicly available.

2. SYLLABLE-BASED RECOGNITION

While triphone-based recognition has for many years been the dominant method of modeling speech acoustics, triphones are a relatively inefficient decompositional unit due to the large number of frequently occurring patterns. Additionally, since a triphone unit spans an extremely short time-interval, such a unit is not suitable for integration of spectral and temporal dependencies. For applications such as SWB, where performance of phone-based approaches is unsatisfactory, focus has shifted to a larger acoustic context. The syllable is one such acoustic unit. Its appeal lies in its close connection to articulation, its integration of some co-articulation phenomena, and its potential for a relatively compact representation of conversational speech.

The use of an acoustic unit with a longer duration also makes it possible to simultaneously exploit temporal and spectral variations. Parameter trajectories [8] and multi-path HMMs [9] are examples of techniques that can exploit the longer acoustic context, but have had marginal impact on triphone-based systems.

	Number of Speakers / Utterances		
	Male	Female	Children
Training	1064 / 24611	1150 / 26405	22 / 500
Dev Test	355 / 8200	384 / 8867	8 / 188
Eval	71 / 1582	77 / 1710	2 / 37

Table 1: A proposed partitioning of the OGI Alphadigit corpus using a 75%/25% partition criterion.

Our recent experiments with syllables on SWB have shown encouraging results, performing on par with comparable triphone systems [2]. Extension of the SWB syllable systems to the alphadigit task was an attempt to validate, on a smaller vocabulary, the approaches taken in our LVCSR system. In addition, the alphadigit task allowed us to isolate phenomena in a domain where lexical problems and pronunciation variation were not dominant.

One problem we are particularly interested in examining is the modeling of monosyllabic words. In standard evaluations on SWB, we found that these words dominated the error rate. Thus, an improvement in monosyllabic word modeling could have a profound effect on the performance of LVCSR systems. The alphadigit task is a good application for evaluating new approaches to monosyllabic word modeling as the alphadigit vocabulary is comprised almost entirely of monosyllabic words.

3. EXPERIMENTS AND RESULTS

We initially developed two baseline systems: a word-internal and a crossword triphone system. Both of these were carefully constructed to provide state-of-the-art performance on a standard SWB task within the constraints of the technology used for implementation. All systems described in this paper were based on a standard LVCSR system developed from a commercially available package — HTK [10]. Again, we framed the baseline alphadigit system around LVCSR technology as our purpose was to validate our SWB results.

3.1. Triphone Systems

Both triphone systems use a phone inventory consisting of 42 phones and a silence model (in addition, a word-level silence model was used). All phone models were standard 3-state left-to-right models without skip states. These models were seeded with a single Gaussian observation distribution. The number of Gaussian mixture components was increased to 32 per state during reestimation using a segmental K-means approach.

A context-dependent phone system was then

bootstrapped from the context-independent system. The triphone models were initialized with a subset of the OGI data consisting of 10% of the training utterances. The single Gaussian monophone models from the context-independent system were clustered and used to seed the triphone models. Four passes of Baum-Welch reestimation were used to generate single-component mixture distributions for the triphone models. These models were then increased to twelve Gaussians per state using a standard divide-by-2 clustering algorithm. The resulting system had 25202 virtual triphones, 3225 real triphones, 9675 states and 12 Gaussians per mixture. The final count for the number of Gaussians is, however, reduced by tying states in the triphones.

3.2. Syllable Systems

The model topology for the syllable models was kept similar to the word-internal phone system. However, each syllable model was allowed to have a unique number of states. Initially, the number of states was determined from our best SWB syllable models. The number of states in these models was selected to be equal to one half the average duration of the syllable, measured in 10 msec. frames. The duration information for these syllables was measured from a forced alignment of SWB data based on a state-of-the-art triphone system. Syllable models were trained in a manner analogous to the word-internal phone system without the clustering stage. The resulting models had 8 Gaussians per state.

3.3. Results and Analysis

Table 2 summarizes the performance of the experiments described in this paper. The context-independent syllable system not only outperforms its triphone counterpart (by approximately 2% absolute), it also outperforms the crossword triphone system by 1% absolute difference. It is also interesting to note the word error rates for both the alphabets and digits separately. The syllable system makes its greatest gains in recognition of the alphabets whereas it lags in performance on the digit recognition.

Table 3 gives an analysis of the primary contributors to error. It is somewhat curious to note that the syllables outperform the triphones in E-set and A-set recognition. One would expect the phones to do better in this arena given their fine-grain phonetic contexts. The phone systems are superior performers on both the nasals and the s-f pairs, however.

Not only do the syllable models achieve a lower word error rate, but they do so in a more efficient manner. Table 4 notes some complexity statistics for both triphone systems and our best syllable system. Notice that the number of models has dropped by an enormous amount from the context-dependent cross-word triphones to the context-independent syllables. The number of total states is a somewhat misleading statistic since the triphone systems use state-tying. Though the syllable system contains more states than

System	Total WER	Alphabet WER	Digit WER
XWRD Triphone	12.2%	15.2%	4.7%
WINT Triphone	13.4%	16.8%	4.8%
Syllable (New Durations)	11.1%	12.8%	6.8%

Table 2: Performance of triphone and syllable systems

Confusion set	Triphone Error Rate	Syllable Error Rate
E-Set	17.7%	16.5%
S-F pair	15.0%	17.6%
A-Set	10.5%	8.3%
Nasals	8.5%	13.2%

Table 3: Percentage of confusions in the respective sets.

System	Logical Models	Real Models	Number of States
XWRD Triphones	25202	3225	2045
WINT Triphones	25202	1011	249
Syllables	42	42	900

Table 4: Complexity of triphone and syllable systems

the word-internal triphones, the search space for the syllable system is significantly smaller. Both of these factors result in a speedup of seven times compared to the triphone system (which means a standard evaluation runs in a day rather than a week).

In the experiments we have done thus far, we have not explored explicit durational modeling in detail. Noise modeling has also not been pursued. To this end, in a recent experiment we defined a new syllable system which uses explicitly marked noise to train the silence model. Thus, the silence model is forced not to discriminate between silence and noise. In earlier systems this was causing problems, because fricatives were being inserted where noise was present in the acoustic data. Additionally, all models in this system were limited to a maximum of 20 states. This is a precursor to an LVCSR experiment using an equal number of states per syllable model.

4. CONCLUSIONS AND FUTURE WORK

Table 2 summarizes the performance of our syllable alphadigit system and comparable triphone systems. The best syllable system performance gives approximately 1% absolute improvement in WER over a comparable crossword triphone system. Though, the syllable system yields performance superior to the triphone system, it falters in manners similar to the triphone systems—namely, the E-set and S-F confusion pair. Thus, the syllable system may be able to gain performance in manners similar to the phone-based systems — context-dependency and state-tying, for example.

Particularly, we are exploring the use of context-dependent (crossword) syllable models. Explicit noise modeling is also being researched since we observed in our analysis that the silence model is generally substituted for fricatives. We have not yet incorporated more sophisticated temporal models into our system. Inclusion of this information could have a significant effect on performance, especially with the S-F pair which has little spectral difference in telephone bandwidth.

5. REFERENCES

[1] P. Loizou and A. Spanias, "High-Performance Alphabet Recognition," *IEEE Trans. on Speech and Audio Proc.*,pp 430-445, Nov. 1996.

[2] A. Ganapathiraju et. al., "Syllable — A Promising Recognition Unit for LVCSR," to be presented at the *1997 IEEE Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, California, USA, December 1997.

[3] S. Euler, B. Juang, C. Lee, and F. Soong, "Statistical segmentation and word modeling techniques in isolated word recognition," in *Proc. of the IEEE ICASSP* '90.

[4] E. Huang and F. Soong, "A probabilistic acoustic MAP based discriminative HMM training," in *Proc. of the IEEE ICASSP* '90.

[5] L. Rabiner and J. Wilpon, "Isolated word recognition using a two-pass pattern recognition approach," in *Proc. of the IEEE ICASSP* '81

[6] http://www.cse.ogi.edu/CSLU/corpora/alphadigit

[7] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," in *Proc. of the IEEE ICASSP '92.*

[8] H. Gish and K. Ng, "Parameter Trajectory Models For Speech Recognition," *Proceedings of the ICSLP*, Philadelphia, Pennsylvania, U.S.A., pp. 466-469, October 1996.

[9] F. Korkmazskiy, et. al., "Generalized Mixture of HMMs for Continuous Speech Recognition", *Proc. of the IEEE ICASSP* '97, pp. 1443-1446, Munich, Germany, April 1997.

[10] P. Woodland, et. al., "HTK Version 1.5: User, Reference and Programmer Manuals", *Cambridge University Engineering Department & Entropic Research Laboratories Inc.*, 1995.

[11] Hamaker, J. et. al., "A Proposal for a Standard Partitioning of the OGI AlphaDigit Corpus," available at *http://isip.msstate.edu/projects/lvcsr/recognition_task/alphadigits/data ogi_alphadigits/trans_eval.text*.