

REDUCED CONTEXT SENSITIVITY IN PERSIAN SPEECH RECOGNITION VIA SYLLABLE MODELING

S.M. Ahadi

Electrical Engineering Department
Amirkabir University of Technology
Tehran, Iran
sma@cic.aku.ac.ir

ABSTRACT: In this paper, an alternative approach to acoustic modeling in Persian continuous speech recognition has been introduced. The approach utilizes syllables in place of phonetic level units due to their more stable and well-defined characteristics. This system is believed to reduce sensitivity to context as part of the context is modeled within syllables. The results obviate the performance of the approach especially in comparison to context-independent models, where reductions up to 21% in the system word error rate, for single-Gaussian case, have been noticed.

1. INTRODUCTION

One of the major issues in implementing automatic speech recognition systems for a language is acoustic modeling. Persian (Farsi) speech recognition has recently been addressed by a number of researchers. Thus, many issues in this regard are to be paid attention. Acoustic modeling is one of these issues. A simple context-independent ASR system for Persian has been introduced in (Ahadi, 1999). Several problems had to be overcome in this system to achieve more acceptable results. One of the main problems in acoustic modeling was the problem of context dependency. Use of the context-dependent acoustic models, such as triphones was a solution for this problem. An alternative is found to be syllable-level modeling.

Syllables, compared to other units of speech, show several desirable specifications. Compared to simple CI units, i.e. monophones, they can model most of the context within a word, while compared to triphones, they are more stable models defined for longer intervals of speech. Furthermore, syllables provide a natural framework for incorporating some prosodic speech features into the recognition process.

Syllables have been used in several efforts for acoustic modeling in different languages. In oriental languages such as Chinese and Korean, syllables are frequently used and modeled due to special syllable-based structures of such languages (Chen *et al.*, 2000) (Kwon, 2000). In other languages, such as English, however, the syllable units are less frequently used as acoustic models due to the large number of such units used in the language, which makes models rather difficult to manage. Furthermore, there exist considerable differences in the size of various syllable units within the language. However, a few approaches to syllabic modeling in English have also been reported, where either the modeling is applied to simpler tasks such as alphan-digit recognition (Hamaker *et al.*, 1998), or sub-syllabic units, such as semi-syllables have been used (Wu *et al.*, 1998).

Although Persian (Farsi) is usually classified as an Indo-European language and its structure resembles more to European languages, rather than oriental languages, it has a much simpler syllabic structure compared to languages like English. This has encouraged us to move to syllable-based speech recognition from our baseline continuous speech recognizer.

The rest of this paper has the following structure: In section 2, the syllabic structure of Persian language is discussed. Our approach to modeling Persian syllables is explained in section 3 and section 4 includes the implementation issues and the results of our experiments. Section 5 concludes this discussion.

2. SYLLABIC STRUCTURE OF PERSIAN

Only three types of syllables are available in Persian, i.e. CV, CVC and CVCC. Moreover, the numbers of vowels and consonants are quite limited (around 7 and 23 respectively). Hence, the total number of possible syllables in Persian, with a simple calculation, is more than 89000. However, there are several other rules governing the structure of a syllable in Persian (Samareh, 1995). These rules seriously limit the actual number of syllables available in the language. As an example, the total number of CVCC syllables used in Persian language is estimated to be 723 and not 85169 possible CVCC combinations (Samareh, 1995).

Table 1 shows the approximate number of different types of syllables encountered in the section of continuous Persian speech corpus, FARSDAT (Bijankhan *et al.*, 1994), used in this research. Note that this is a medium-sized vocabulary speech corpus with about 1100 word entries and is so far the only available multi-speaker continuous Persian speech corpus. Due to limited size of the vocabulary, the number of syllables is far less than the total number of syllables which are encountered in practice and consists of 842 syllables in total. Although every effort was carried out to extract all pronounced syllables, since in some cases, some words were pronounced differently by various speakers, different types of syllables would have resulted which might not have been spotted in all cases. Hence these results may be considered as approximate.

Syllable type	CV	CVC	CVCC	Total
# in database	146	531	165	842

Table 1. Approximate number of different types of syllables available in our speech corpus.

3. SYLLABLES AS ACOUSTIC MODELS

In our context-independent speech recognizer (Ahadi, 1999), 32 models were defined for 30 basic Persian phonemes plus two silence and between-word space models. Most of these models were 3-state left-right HMMs with no skip transitions. The only exception to this was the model used for between-word space, which consisted of only one state with the possibility of being bypassed.

Our first approach to building syllable models consisted of concatenating previously trained context-independent models. This simple approach was followed due to this fact that well trained initial monophone models can increase the possibility of better model training with relatively limited training data. In fact this is a shortcut to syllable model training, while otherwise, the training would have been very cumbersome. This led to 6-state to 12-state HMMs according to the type of the syllable being modeled. The constructed models were then further trained using the available training data.

3.1 Syllable Models with Reduced Number of States

Although the above approach resulted in better modeling of speech, it also increased the number of system parameters substantially. In fact, not only the number of models increased from 32 to 844 (including silence models), but also the number of states per model increased from 3 to 6, 9 or 12. Hence, the system was not only unsuitable for use with mixture Gaussians due to the relatively limited amount of training data available, but its ability to perform well in the single-Gaussian modeling case was also under question.

A first solution for this problem was to reduce the number of states per model for syllable models. This was carried out by reducing the number of states per syllable model to 4, 5 and 6 states for CV, CVC and CVCC models respectively. Obviously in this approach, the models could not be constructed by concatenating the monophone models, as the number of states per phone in the syllable model did not correspond directly to the number of states of the monophone models. The implementation issues regarding this type of models will be discussed in section 4.

3.2 Syllable Models with tied parameters

In order to further reduce the number of total system parameters, a tying approach was followed. In this phase the state parameters of the trained models of our Initial system, i.e. those with 6 to 12 states per model, were used. The process consisted of clustering the corresponding states of all similar phones within the syllable models and tying the parameters of the states placed in the same cluster. The clustering algorithm was an agglomerative one (Everitt, 1993) and can be described as follows:

1. Allocate one cluster per state.
2. Find all inter-cluster distances.
3. Find the smallest inter-cluster distance ($d(i,j)$).
4. If $d(i,j)$ is not less than T , stop.
5. Otherwise, merge clusters i and j and find all inter-cluster distances with this cluster.
6. Continue from 3.

In the above algorithm, T is a predefined inter-cluster threshold. The distances between states were calculated using divergence distance metric. This algorithm was applied to all sets of states chosen as explained and continued until converged. In the end, the state parameters of all states in the same cluster were tied and another phase of training was carried out.

4. IMPLEMENTATION AND RESULTS

As explained, The speech corpus used in these experiments was FARSDAT. Only part of the database, containing speech data from fluent Persian speakers was used. This section of the corpus consisted of utterances from 137 speakers. These speakers were then divided into 2 parts to form the training and test sections of the database, containing around 1800 and 900 sentences respectively, with a vocabulary of about 1100 words.

The speech data was then downsampled from 44.1 KHz to 16 KHz and the parameterization phase was performed by applying a pre-emphasis filter with a coefficient of 0.95 and blocking into frames of 25 msec.

with 15 msec. of overlap. A hamming window was also applied to the signal to reduce the effects of frame edge discontinuities. 12 Mel-cepstral coefficients plus log energy were then computed and the delta and delta-delta parameters added to extend this number to 39. This parameterized data was then used throughout all training and test procedures.

Our earlier experiments showed that applying even a simple language model during the recognition process would unexpectedly increase the system performance. This was found to be according to the structure of the speech corpus and its relatively low grammar perplexity, since it was not specifically designed for speech recognition purposes. Hence in this paper, the reported results are from the tests that did not utilize any type of language model.

The initial implementation stage used the previously trained context-independent models as explained in section 3. The concatenated syllable models were then trained iteratively with all the available training sentences, using syllable-based transcriptions. The resultant system's performance in comparison to our baseline monophone system indicated around 10% reduction in the word error rate (from 53.1% for the baseline to 43.3%). However, this system's high number of parameters did not allow appropriate training with larger numbers of mixture components.

In the second approach, to build reduced-state syllable models, an approach similar to the previous case was followed. However, due to unavailability of monophone models with smaller state count, such models had to be built initially. This was done by training single and 2-state monophone models using a procedure similar to that used for building the baseline system. The syllable models were then constructed by concatenating these models using single state models for the phones appearing as middle phones in the syllable and 2 state models for those appearing at the start or end positions in the syllable. The resultant syllable models were considered as initial models for training. Individual model training was the next phase. However, due to the large number of syllable models, small number of available time-labeled training sentences (only 119 such sentences were available) did not suffice. Hence, time-labeled transcriptions for about 600 of the training sentences were generated and used for this purpose. The available training data for many syllables was considered insufficient. This was due to this fact that about 264 syllables, i.e. about 31% of all the syllables, appeared less than 3 times among training data. MAP estimation was used for further training of the models in order to overcome the problem of sparseness of the training data. The resultant system, performed only slightly worse than previous system (43.8% word error rate), but a reduction of about 45% in the number of the system parameters was obtained. Although this was a noticeable reduction in the system parameter count, as will be mentioned later, the mixture Gaussian results did not improve much using the latter system. The system parameters seemed to be still too high to allow mixture-Gaussian modeling improve the performance.

System	Initial	Reduced-state	Tied-parameter
Total No. of states	7639	4233	1018
Reduction in parameter count	-	44.5%	86.7%

Table 2. Comparison of the total Number of states in 3 implemented systems and percentage of reduction in the number of parameters in the reduced-state and tied-parameter systems in comparison to our initial system.

The third system with tied parameters was introduced to overcome this problem. In this approach, larger number of parameters were allowed to appear in the system, as a tying procedure was going to be applied. Hence, the initial system which had performed slightly better was utilized here. The approach explained in section 3.1 was implemented using our already trained initial system, with the threshold levels

experimentally set. The number of parameters in the resultant system, in comparison to the number of parameters in our 2 other systems, is shown in Table 2.

The large amount of reduction in the system parameter count of the tied-state system not only resulted in a better system for mixture-Gaussian modeling, but also increased the recognition accuracy of the single-Gaussian case in comparison to 2 other systems. Table 3 compares the performance of our 3 syllable-based systems to that of our baseline monophone system. This table also includes results for two triphone context-dependent systems implemented for comparison purposes. The first of these two systems was developed initially and the second, which is a tied-parameter triphone system, was implemented using the first triphone system and an approach similar to that used for tying in our syllable-based system. The total number of states in the latter case was 1026, which is very close to the total number of states in our tied-parameter syllable-based system. All the systems indicated used single-Gaussian pdfs. Large improvements obtained by syllable modeling, in comparison to the baseline system, is obvious. Parameter tying has also contributed significantly to further improvements in the results. Moreover, the syllable-based systems have marginally outperformed the triphone systems in both tied and untied cases.

Assessment of the performance of syllable based mixture-Gaussian models was also carried out. This phase was implemented for all above-mentioned syllable-based systems. However, as expected, the test results with initial and reduced-state syllable-based model sets did not improve much with multiple components. Further increase in the number of mixture components to 3 or more even deteriorated the results. This was because of the large number of parameters in the model sets that resulted in unreliable training. However, the results with the mixture-Gaussian tied-parameter system were satisfactory. These results are shown in Table 4.

System	Baseline	Context-dependent	Tied-parameter Context-dependent	Initial Syllable-based	Reduced-state Syllable-based	Tied-parameter Syllable-based
Percent word error rate	53.1	43.9	32.2	43.3	43.8	31.7
WER reduction compared to baseline	-	9.2	20.9	9.8	9.3	21.4

Table 3. Recognition results of the 3 syllable-based systems in comparison to the baseline monophone and two context-dependent triphone system.

Number of mixture components	1	2	3	5
Word error rate	31.7	24.9	23.0	21.9

Table 4. Mixture-Gaussian syllable-based tied-parameter system performance.

5. CONCLUSION

In this paper, the issue of syllable-based acoustic modeling for Persian continuous speech recognition was discussed. It has been shown that the syllable models can perform well in Persian continuous speech recognition tasks due to special characteristics of the language. However, their number tends to increase in larger vocabularies and a tying approach can be useful. The tied-parameter system has performed well, even in comparison to context-dependent systems, making mixture-Gaussian modeling with relatively limited available training data possible. There are hopes that the application of this approach to large vocabulary continuous speech recognizers will also be useful in reducing the overall word error rates.

6. REFERENCES

- Ahadi, S.M. (1999) "Recognition of continuous Persian speech using a medium-sized Vocabulary speech corpus", Proc. EUROSpeech, vol.2, pp. 863-866.
- Bijankhan M. *et al.* (1994) "The speech database of Farsi spoken language", Proc. 5th Australian International conference on Speech Science and Technology (SST).
- Chen, B., Wang, H. & Lee, L. (2000) "Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics", Proc. International Conference on Acoustics, Speech and Signal Processing.
- Everitt, B.S. (1993) Cluster Analysis, 3rd edition, (Edward Arnold: London).
- Hamaker, J., Ganapathiraju, A., Picone, J. & Godfrey, J.J. (1998) "Advances in alphadigit recognition using syllables", Proc. International Conference on Acoustics, Speech and Signal Processing, pp. 421-424.
- Kwon, O-W. (2000) "Performance of LVCSR with morpheme-based and syllable-based recognition units", Proc. International Conference on Acoustics, Speech and Signal Processing.
- Samareh, Y. (1995) Persian language phonetics, 4th edition, (University Publications Center: Tehran), (In Persian).
- Wu, S-L., Kingsbury, B.E.D., Morgan, N. & Greenberg, S. (1998) "Incorporating information from syllable-length time scales into automatic speech recognition", Proc. International Conference on Acoustics, Speech and Signal Processing, pp. 721-724.