AN EXPERIMENTAL MULTI-SPEAKER STUDY ON FARSI PHONEME DURATION RULES USING AUTOMATIC ALIGNMENT

D. Gharavian, H. Sheikhzadeh, and S. M. Ahadi
Department of Electrical Eng., Amirkabir University of Technology (Tehran Polytechnic)
Tehran,Iran
Emails: hsheikh@cic.aku.ac.ir and sma@cic.aku.ac.ir

ABSTRACT: In this paper we present the results of an experimental study on phoneme duration rules of the Farsi (Persian) language. A multi-speaker speech corpus is used and an automatic alignment algorithm based on CD-HMMs is employed. The results are utilized to examine a few duration rules stated in the literature. After completion of this research, the results could be employed in the frameworks of speech synthesis and speech recognition.

1. INTRODUCTION

There have been few scientific works to explore Farsi phoneme duration rules. While for a long period of time, linguists have been speculating on phoneme duration rules of the language, none of the works have been experimentally validated (Samareh, Y., 1995). Therefore, one sometimes finds conflicting results in the existing literature. A reasonable knowledge of the duration rules would be beneficial (and sometimes necessary) to automatic speech recognition, speech synthesis and prosodic modeling. For example in our rule-based (Sheikhzadeh, H. *et.al* , 1999) and Multi-Band Resynthesis (MBR) speech synthesizers, it is crucial to implement the phoneme duration rules to obtain intelligible speech. Also in speech labeling and recognition, a good knowledge of the phoneme duration patterns would help to decrease labeling and classification errors, respectively. Finally, from a pure linguistic point of view, it is interesting and helpful to know more about the actual duration rules of the uttered phonemes. In this research, we have employed a Farsi corpus known as FarsDat (Bijankhan M. *et al.*, 1994) to experimentally uncover some of the duration rules of the Farsi language. At this stage, we have limited our attention mostly to phoneme durations and particularly to vowels. The automatic labeling has been done by using phoneme-based CD-HMMs. In this research, the following cases have been explored:

1. The duration of tense vowels compared to short vowels. In Farsi, there are three tense vowels of /a/, /i/, and /u/, and correspondingly three short vowels of /æ/, /ɛ/, and /o/. Currently, there is no concrete knowledge available about the duration ratios of tense to short vowels at different phonetic and syllabic contexts.

2. The duration of similar vowels at different Farsi syllabic contexts of CV, CVC, and CVCC. It is claimed that vowels are longer in CVCC syllables than similar vowels in CVC or CV syllables.

3. The duration of similar vowels at various phonetic contexts. For example, it is claimed that a vowel in a CVC syllable ending in nasal /m/ would be longer in duration than the same vowel in a CVC syllable ending in nasal /n/.

4. The average duration of various Farsi phonemes. After reporting the results, a few conclusions are made.

This paper is organized as follows. First the speech corpus and the automatic alignment procedure are briefly introduced. Next the duration results are presented, first for vowel durations and then for all Farsi phonemes.

2. AUTOMATIC ALIGNMENT

The speech corpus used for this work was FarsDat (Bijankhan M. *et al.*, 1994). The number of sentences from this corpus, used for training, was about 1814, uttered by 137 male and female speakers. The parameterization was carried out by extracting 12 MFCC and log energy parameters and their delta and delta-delta coefficients. The alignment procedure consisted of Viterbi forced

alignment using phoneme-based models. The number of Farsi phonemes used in this research was 30, as shown in Table 1 (IPA93 phonetic representation is used throughout this paper). The HMMs used for this purpose consisted of 3-state models for all Farsi phonemes and silence, and a 1-state model for between-word space. These monophone models were initialized using 119 time-aligned training sentences and further training was carried out using Baum-Welch algorithm with the same sentences. Later, another Baum-Welch training phase was carried out using all 1814 sentences and their FSNs (Finite state Networks) (Rabiner, L. & Juang, B.H., 1993) (Ahadi, S.M., 1999). Two sets of trained models were built in this manner, one with 7-Gaussians and another with 15-Gaussians per state. The latter showed a performance of about 78% word recognition accuracy during recognition. Both the above systems were tested during alignment procedure and the difference noticed was negligible. So, the 15-Gaussian model set was used during our investigations. The alignment results, compared to hand labeling, did not show any noticeable difference.

Phonetic Groups	Phoneme 1	Phoneme 2	Phoneme 3	Phoneme 4	Phoneme 5	Phoneme 6	Phoneme 7	Phoneme 8
Vowels	/ɑ/ xɑb	/æ/ sæbr	/ɛ/ tʃɛrɑɢ	/u/ ruz	/i/ diruz	/o/ gozæʃt	/ow/ rowʃæn	----------
Liquids	/r/ rɑh	/l/ lɑlɛ	----------	----------	----------	----------	----------	----------
Glide	/j/ mɛjdɑn	----------	----------	----------	----------	----------	----------	----------
Nasals	/m/ mærdom	/n/ nɑmɛ	----------	----------	----------	----------	----------	----------
Plosives	/b/ bɑzɑr	/p/ pɑrɛ	/t/ tærtib	/d/ dæʃt	/ɢ/ ɢævi	/k/ kɛtɑb	/g/ goruh	/ʔ/ mæʔlum
Fricatives	/f/ fɑrsi	/s/ særmɑ	/v/ vɑrzɛʃ	/x/ xɑnɛ	/z/ zɛræng	/ʒ/ ʒɑlɛ	/ʃ/ ʃoʔlɛ	/h/ huʃ
Affricates	/dʒ/ dʒɑmɛd	/tʃ/ tʃirɛ	----------	----------	----------	----------	----------	----------

Table 1: List of Farsi phonemes.

2.1 Speech Rate Normalization and Statistical Processing

To normalize the duration results for the speech rates of various speakers, the adopted method was to normalize all duration results so that all speakers had similar average vowel durations.

To have statistically valid results, we further processed the durations obtained through automatic alignment by four different methods:

Method 1: When reporting absolute phoneme durations, we employed a variance analysis of the duration results. Those values that differed from the mean value by one standard deviation were discarded and the mean was re-calculated.

Method 2: When reporting duration ratios (like phoneme /a/ to /æ/ duration ratio), one approach was to first apply method 1 to each of the two duration sets for all speakers, and then find the ratio of the mean values.

Method 3: The other method of reporting duration ratios was to first derive the average duration ratios of the two phonemes involved for each speaker separately, and then average the ratios across different speakers.

Method 4: After using method 3, the variance analysis (method 1) was applied to the ratios.

When the durations were all normalized for speaker rates, methods 2, 3, and 4 yielded very similar results and thus we mostly reported the ratio results obtained by method 4 only, unless otherwise stated. In some cases where we did not have enough syllables (specially of CVCC type) from all speakers, we employed method 2 to avoid too much data skipping. In reporting the results of such tests, out of the two fractions of standard deviation to mean (for the two phonemes), the maximum value was reported (Table 3, Table 5, and Table 8).

3. VOWEL DURATIONS

3.1 Long Versus Short Vowels

Table 2 shows the duration ratios of the three Farsi tense vowels to their (acoustically) closest short vowels. As shown, the ratios of standard deviation to mean values are small. Since the vowel duration in Farsi depends on the syllable type containing the vowel, the duration ratios were separately processed for the three syllable types of the language. Method 2 was employed here to avoid too much data skipping. The results presented in Table 3 (dashed line in table means that there was not enough data available) show that, as speculated, the ratios differ not only for various phoneme pairs but also for different syllable types. Presented in Table 4 is the average duration of Farsi vowels for male and female speakers. As expected, the results are very close for both genders, however it seems that female speakers have a longer /æ/ phoneme.

Duration ratio	Mean	Std/Mean
/a/ to /æ/	1.1698	0.1193
/i/ to /ɛ/	1.6357	0.0983
/u/ to /o/	1.4010	0.1477

Table 2: Ratio of tense to short vowels for all syllables.

Vowels	CV syllables		CVC syllables		CVCC syllables	
Duration ratio	Mean	Max. Std/Mean	Mean	Max. Std/Mean	Mean	Max. Std/Mean
/a/ to /æ/	1.3445	0.1796	1.3229	0.1349	1.1892	0.1289
/i/ to /ɛ/	1.6299	0.1536	1.3913	0.1606	1.3441	0.1757
/u/ to /o/	1.5346	0.2131	1.2895	0.1848	----	----

Table 3: Ratio of tense to short vowels for CV, CVC and CVCC syllables.

Vowel	Male Speaker		Female Speaker	
Duration	Mean (ms)	Std/Mean	Mean (ms)	Std/Mean
/æ/	82.64	0.0678	97.96	0.0629
/a/	102.54	0.0538	99.53	0.0475
/ɛ/	59.58	0.0666	63.66	0.0540
/i/	99.75	0.0510	100.16	0.0583
/o/	72.27	0.0710	75.05	0.0539
/u/	104.89	0.0787	97.20	0.0819

Table 4: Duration of vowels in male and female speakers.

3.2 Vowels in Different Syllabic Types

There are many speculations about the duration of vowels in different Farsi syllables. As Table 5 and Table 6 show, vowels in CVCC syllables are longer than their similars in CVC and CV syllables. Generally speaking, vowels are longer in duration in syllables containing more consonant clusters. However, the ratios are totally dependent on the vowel itself. For example, in case of vowel /i/, the duration ratio for CVC to CV is close to one.

Vowels	CVC / CV		CVCC / CV		CVCC / CV	
Duration ratio	Mean	Max. Std/Mean	Mean	Max. Std/Mean	Mean	Max. Std/Mean
/æ/	1.3740	0.1014	1.8120	0.1285	1.3188	0.1285
/a/	1.3519	0.1796	1.6027	0.1796	1.1855	0.1349
/ɛ/	1.2755	0.1406	2.1675	0.1563	1.6994	0.1563
/i/	1.0888	0.1606	1.7874	0.1757	1.6416	0.1757
/o/	1.4385	0.1310	2.5008	0.1817	1.7385	0.1817
/u/	1.2087	0.2131	1.2702	0.2131	1.0509	0.1848

Table 5: Duration ratio of vowels in different syllabic contexts.

Vowels	CV Syllable		CVC Syllable		CVCC Syllables	
Duration	Mean (ms)	Std/Mean	Mean (ms)	Std/Mean	Mean (ms)	Std/Mean
/æ/	68.20	0.0960	93.71	0.1014	123.59	0.1285
/a/	91.70	0.1796	123.98	0.1349	146.97	0.1289
/ɛ/	55.03	0.1406	70.19	0.1179	119.28	0.1563
/i/	89.70	0.1536	97.66	0.1606	160.32	0.1757
/o/	55.15	0.1202	79.33	0.1310	137.92	0.1817
/u/	84.63	0.2131	102.30	0.1848	107.50	0.1641

Table 6: Vowel durations in different syllabic contexts.

3.3 Vowels Followed by Different Consonants

A rule stated in (Samareh, Y., 1995) is that vowels have a longer duration when followed by /m/ rather than /n/. A second rule is that vowels are longer if followed by voiced consonants rather than unvoiced consonants. We experimentally explored these rules. As demonstrated in Table 7, the first rule is mostly true for syllables ending in /m/ and /n/ nasals. However, it is not true in the case of vowels /ɛ/ and /o/. The second duration rule is less true as shown in the table. We believe that less general duration rules have to be investigated.

Vowels	/m/ to /n/		Voiced to Unvoiced	
Duration Ratio	Mean	Std/Mean	Mean	Std/Mean
/æ/	1.1393	0.1394	1.0434	0.0904
/a/	1.1457	0.1673	1.3072	0.0970
/ɛ/	0.8722	0.2023	1.0600	0.1582
/i/	1.2642	0.1958	0.7451	0.1768
/o/	0.9885	0.2095	1.0822	0.1929
/u/	1.5033	0.1889	1.1787	0.1819

Table 7: Duration ratio of vowels in CVC syllables ending in /m/ to those ending in /n/, also ending in voice consonants to those ending in unvoiced consonants.

4. DURATION OF DIFFERENT FARSI PHONEMES

Shown in Figure 1 are average durations of all Farsi phonemes. The results are ordered such that phonemes in the same phonetic classes follow each other. We have also explored a few cases of consonant durations. As Table 8 shows, in most cases, an unvoiced consonant is longer than its voiced phonetic pair. As the figure shows, liguids are the shortest in duration, while unvoiced fricatives are the longest.

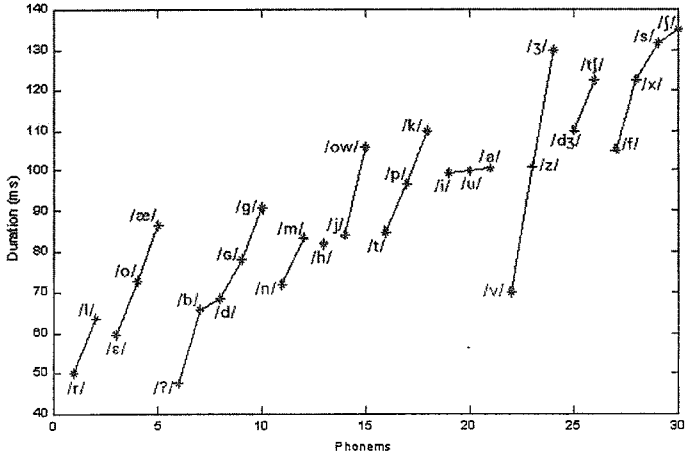

Figure 1 : Average duration of all Farsi phonemes.

Phoneme	Average ratio		First consonant in CVC syllables		Second consonant in CVC syllables	
Duration Ratio	Mean	Std/Mean	Mean	Max. Std/Mean	Mean	Max. Std/Mean
/f/ to /v/	1.4981	0.1117	1.3859	0.1667	1.3480	0.1712
/ʃ/ to /ʒ/	1.0397	0.0863	1.1567	0.1704	0.9533	0.1494
/p/ to /b/	1.4696	0.1505	1.4527	0.1289	1.2332	0.1578
/t/ to /d/	1.2491	0.0817	1.5950	0.1032	1.3162	0.1504
/k/ to /g/	1.2181	0.0915	1.2332	0.1069	1.0135	0.1656
/tʃ/ to /dʒ/	1.1176	0.1216	1.0732	0.1322	1.1496	0.1796

Table 8: Duration ratios of phoneme pairs.

5. CONCLUSION

This is a first attempt to experimentally validate the rules governing phoneme duration in Farsi language. More research has to be done to complete this work before one can actually use the results in a speech engineering application. We are now in the process of exploring more detailed rules for phonemes, and expanding the work to include syllables and words. Also, a detailed study of energy and pitch patterns would be the future extension of this research.

6. REFERENCES

Ahadi, S.M. (1999) "Recognition of continuos persian speech using a medium-sized vocabulary speech corpus", Proc. EUROSPEECH-99, 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).

Rabiner, L. & Juang, B.H. (1993) Fundamentals of speech recognition, (Prentice Hall).

Samareh, Y. (1995) Persian language phonetics, 4th edition, (University Publications Center: Tehran), (In Persian).

Sheikhzadeh, H. *et.al* (1999) "Farsi Language Prosodic Structure, Research and Implementation Using a Speech Synthesizer", Proceedings of Eurospeech, Budapest, Vol. 4, pp. 1647-150.