Speaker-specific variability of Phoneme Durations

Charl J. van Heerden[1], *Etienne Barnard*[2]

[1]Department of Electrical, Electronic and Computer Engineering, University of Pretoria, South Africa
[2]School of Electrical, Electronic and Computer Engineering, University of North-West, South Africa
[1,2]Human Language Technologies Group, Meraka Institute, South Africa

cvheerden@csir.co.za, ebarnard@csir.co.za

Abstract

The durations of phonemes varies for different speakers. To this end, the correlations between phonemes across different speakers are studied and a novel approach to predict unknown phoneme durations from the values of known phoneme durations for a particular speaker are presented, based on the maximum likelihood criterion. Several interesting patterns are observed. Phonemes from the same broad phonetic class tend to covary most strongly (and therefore intra-class predictions of unknown phoneme durations are most accurate), but significant cross-class correlations are also present. Consequently, knowledge of only a few highly-correlated phonemes' durations is necessary to make a good duration prediction.

Index Terms: phoneme durations, speech recognition, maximum likelihood, eigen vectors

1. Introduction

Developing accurate phoneme duration models has been a topic of discussion for several years, especially with regard to the potential benefits for automatic speech recognition (ASR) [1]. In [2],[3] we showed that accurate phoneme duration models can significantly improve state of the art speaker recognition (SR) systems in a text-dependent environment. For practical applications of both ASR and speaker recognition, duration models have to be developed for text-independent speech. This is not a trivial problem as there are many factors influencing the duration of phonemes in text-independent speech, such as position in word, position in sentence, stress, preceding and following phonemes, speech rate etc. Although the work done in [2] was in a text-dependent environment, it did confirm earlier findings by [4] that phoneme durations are also speaker-specific to a large extent, which adds another dimension to the model estimation. All these factors contribute to making data scarcity a significant obstacle to characterizing phoneme durations accurately. This obstacle, which was first identified in 1988 already by Crystal and House [5], remains arguably the most significant one to the more general use of phoneme durations.

An attempt to estimate the individual contributions of the abovementioned factors to the total variance was made by [1]. A hierarchical analysis of variance was performed and it was found that much of the variance can indeed be explained by these factors. Because of the type of ANOVA performed, it was not possible to examine interactions among the factors, which may omit important information. Duration patterns were also modelled by [6],[7],[8],[9] in order to improve speaker recognition performance. It was observed that significant improvements in accuracy can be achieved by separately modelling word durations, single phoneme durations and state durations using 3-state hidden Markov models (HMMs). Data sparseness was addressed in all cases by a back-off technique, though which word-models would be backed off to triphone models and the latter to single phoneme models. This ignores the effect of the specific factor being addressed on the particular phoneme. Rao Gadde [8] also performed a simple speech rate normalization. The speech rate was calculated as the number of phonemes per second. By applying this simple normalization technique, a consistent improvement in word recognition was observed over several databases.

Taken together, these studies are strong evidence that accurate phoneme duration models can greatly benefit both ASR and speaker recognition. However, no sophisticated model exists yet because of data scarcity (which limits the number of factors that can be modeled), the many different factors which have an influence on the duration of phonemes and the fact that interaction effects between the different factors are not incorporated into the models.

In this paper we present some introductory work towards the goal of building a model that accurately incorporates all of the abovementioned factors and their interactions. In particular, we have focused on two of the factors that have been found to be important, namely "speaker", and "phoneme type/class". Our objective was to see if it is possible to make better duration predictions of unknown speakers than the back-off approach, given a model that was trained from other speakers' data. We also believe that certain phonemes are more predictable than others and that certain classes of phonemes tend to have greater influence on the durations of others. All of these experiments were conducted on the TIMIT corpus because of the availability of accurate manual phoneme segmentations.

2. TIMIT corpus

The TIMIT corpus is a speech corpus of 630 speakers from eight major dialect regions in the United States. Each speaker spoke 10 utterances resulting in 6300 utterances in TIMIT. The training set consists of 462 speakers, which comprise 326 males and 136 females. Three types of sentences were read: sx, si and sa. The sx sentences were read from a list of 450 phonetically balanced sentences that were designed at MIT, the si sentences from 1890 phonetically diverse sentences designed at TI and two dialect sentences designed at SRI. The test set consists of 168 speakers, which were selected so that no sentence text appears in both the training and test set.

3. Modelling approach

The modelling approach we took was influenced by two questions. From a theoretical viewpoint we wanted to know if and how different phonemes' durations covary under different conditions such as those mentioned above. In particular we decided to investigate this question under the "speaker" condition. In practical terms, we wanted to see whether it is possible to reduce the data requirements of phoneme duration predictions by using inter-phoneme information. This knowledge would be useful for scenarios where data scarcity is an issue.

As already mentioned, there are several factors that act together in a complex and as yet unknown fashion in influencing the durations of the phonemes. A good understanding of each factor is necessary before attempting to model them together. In answering the questions we posed, we wanted to isolate the "speaker" factor. For that reason, we decided to conduct independent experiments where we worked with mean phoneme durations per speaker in an attempt to smooth out the other factors. Our first set of measurements therefore consisted of computing the correlations between mean phoneme durations across speakers.

In order to get a perspective of the extent of the influence of factor on the durations, an eigenvector analysis was done. The directions and magnitudes of the principal contributions to variance were obtained by calculating the eigenvalues and eigenvectors. By then projecting speaker-specific data onto the eigenvectors a good indication is obtained of the speaker differences for the specific factor. The directions of the eigenvectors explain how each of the input factors contributes to the specific dimension.

The eigenvectors and eigenvalues are obtained from the covariance matrix of the $n \times m$ data where n is the number of levels of a factor (number of speakers in our case) and m the number of phonemes. Before calculating the covariance matrix, the data matrix is normalized by subtracting the mean values from the column vectors and then dividing by the standard deviation. This ensures that phonemes with a high variance do not dominate the analysis.

We decided to use a maximum likelihood (ML) approach for cross-phoneme duration estimation, because this enabled us to utilize the information provided by the eigenvectors in a practical model. It was assumed that the data can be approximated by a normal distribution with the covariance matrix calculated as described above. Suppose one has a vector $\overline{x} = \{x_0 ... x_m\}$ representing normalized phoneme durations with x_{m-p} unknown, $p < m$. The ML approach will find x_{m-p} such that the probability $P(x_0 ... x_m)$ is maximized. If one defines x to be $\overline{x} = \overline{d} - \overline{mu}$ with d the original duration, the ML solution given $\Sigma^{-1}\overline{x}$ can be found from

$$\frac{\partial \Sigma}{\partial x_{m-p}} = 0 \tag{1}$$

The solution to (1) is simply

$$\Sigma^{-1}\overline{x} = 0 \tag{2}$$

on condition that $\overline{x} = \overline{k}$ with the exception of x_{m-p}, with $\overline{k}$ being the given data vector. (2) can easily be solved by simple linear algebra. This method was then extended by allowing several unknown durations to be estimated simultaneously using exactly the same approach as described above.

4. Experimental setup

The proposed models were tested using the TIMIT database as described in section 2. TIMIT contains 52 different phone symbols. This set was reduced to the well-known ARPABET owing to data scarcity, which consists of 48 symbols, by combining em, en and eng into en, hh and hv to h and zh and z to z. ARPABET was then reduced by one symbol to 47 symbols by combining [ʊ] and [ʌ].

The training set of 462 speakers was used to estimate the covariance matrix, as well as the eigenvectors and eigenvalues. The test set of 168 speakers was then used to test the models by trying to predict phoneme durations. For every speaker, sample means of all present phonemes were calculated.

Our initial measurements of the correlations in phoneme durations across speakers focused on the relationships between phonemes that covary in duration. Thereafter, we conducted a number of experiments to investigate duration prediction using the ML method described above.

4.1. Experiment 1

An iterative approach was used to estimate every single phoneme duration, given all the other phoneme durations. The objective of this experiment was twofold: to determine if the ML approach can be used to make better predictions than simply predicting the global mean of each phoneme and also to determine the relative predictability of individual phonemes.

4.2. Experiment 2

The same iterative approach was used to estimate the durations of all phonemes, but this time only phonemes of the same class as the phoneme in question were given as input. The hypothesis that was to be tested was that phonemes tend to vary in classes. There were five phoneme classes: stops, fricatives & affricates, nasals, semivowels & glides and vowels.

4.3. Experiment 3

Experiment 2 was repeated, but instead of using phoneme durations of the same class, all durations *except* the durations of phones of the same class were given as input. An interesting observation from the covariance matrix was tested here in that there seems to be a "cross-class" correlation between certain phonemes.

4.4. Experiment 4

For every speaker 50 duration estimates were done. Every estimation entailed three vowels and three consonants to be estimated simultaneously, with the rest of the observed phonemes given as data to the ML model estimator.

4.5. Experiment 5

For every speaker, each phoneme was estimated iteratively, each time adding phonemes in descending order according to their correlation with the phoneme to be estimated.

4.6. Experiment 6

The theoretical minimum of experiment 5 was calculated by adding phonemes until just before the error started to increase again.

5. Results

5.1. The correlation of phoneme durations

The correlations across speakers between the durations of all phonemes were computed; because of the normalization em-

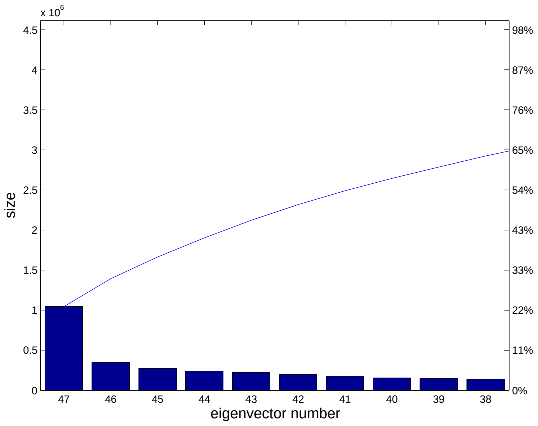

Figure 1: *Pareto chart of the eigenvalues obtained from the speaker/phoneme covariance matrix.*

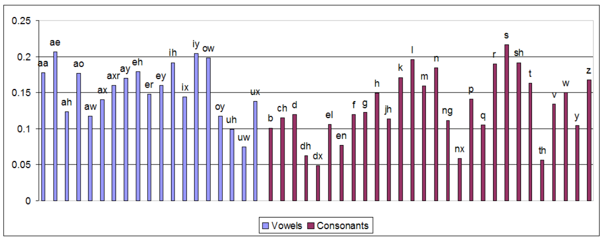

Figure 2: *Components of first eigenvector.*

ployed, these are equivalent to Pearson correlation coefficients. Some typical results are shown in figures 11 to 14, which represent the largest and smallest measured correlation values between four different phonemes and all other phonemes in our set. We see that some groups of phonemes (including most vowels) have high correlations with all other phonemes in the same group. Other phonemes have a more diverse set of correlations - for example, the duration of “p” correlates highly not only with other plosives (“k” and “t” in Figure 12), but also with the fricatives “s”, “z” and “sh”, the nasal “n”, etc. Similarly, the duration of “r” in Figure 13 correlates highly with the expected “l” and “w”, but also with several vowels. Finally, some phonemes have few strong correlations - for example, “dx” in Figure 14, which has reasonably weak correlation with the other flap (“nx”), and no other notable correlations.

5.2. Eigenvector analysis

We now describe the results obtained in our analysis of the eigenstructure of the correlation matrix. Firstly, the magnitudes of the eigenvalues indicate how much weight or value a particular eigenvector carries.

From Fig. 1 it can be seen that the first eigenvector contains more than 22% of the total information and that approximately 65% of the information is contained within the first 10 eigenvectors. This is a strong indication that a significant amount of information is contained in a relatively small number of factors.

As can be seen from Fig. 2, the first eigenvector corresponds to a simultaneous stretching of *all* phonemes - this can therefore be seen as an indication of speaking rate. The vowels and fricatives are seen to be the most consistent participants in this change. The second eigenvector, shown in Fig. 3, corresponds to a differential lengthening of vowels in comparison with consonants, whereas the third eigenvector (Fig. 4) seems to indicate a distinction between the relative lengths of liquids, glides and nasals, on the one hand, in comparison with plosives, fricatives and certain vowels, on the other.

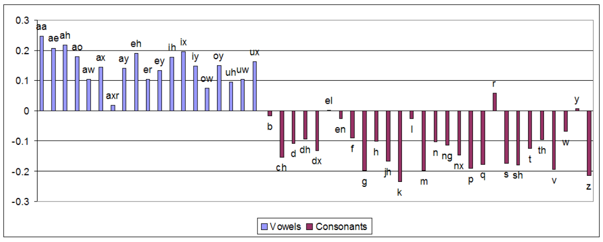

Figure 3: *Components of second eigenvector.*

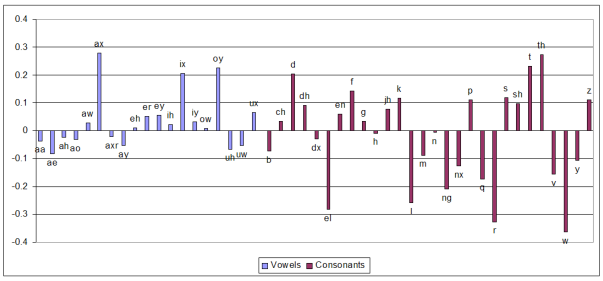

Figure 4: *Components of third eigenvector.*

5.3. ML analysis

The performance of the ML model in the five experiments described in Section 4 was calculated in terms of the variance normalized mean squared error (MSE) between the correct duration and the estimated duration. A total of 7522 estimations were done for the first three and also the last experiment and 50400 for the fourth. The latter will be normalized to the other four experiments in order to give comparable results. A baseline against which the results can be tested must also be established. Two baselines were selected: a nearest neighbor approach (where the closest training speaker based on all the known phoneme durations is calculated, using the Euclidean distance) and simply using the global mean for the specific phoneme. The results can be seen in Table 1. Experiment 5 was conducted to evaluate individual phoneme errors and is thus not presented in the table.

Table 1: *Variance normalized MSE of the ML model, nearest neighbor and mean model from the four experiments.*

Exp.	ML	Global Mean	Eucl. dist
1	0.874	1.070	1.601
2	0.871	1.070	1.658
3	1.007	1.070	1.730
4	0.874	1.087	1.606
6	0.815		

Several interesting observations can be made from Table 1. Firstly, we note that the ML approach consistently outperforms the global mean approach. In experiment 1 the percentage improvement is approximately 18.3% and this increases to 18.6% for experiment 2. As could be expected, this percentage drops significantly for experiment 3 (to only 5.9%). The interesting observation here is that this approach still performs better than the global mean approach. This phenomenon confirms that the many non-zero correlations between different classes of phonemes can be employed usefully. Surprisingly, the average improvement jumps to 19.6% for experiment 4 where six unknown phoneme durations were estimated simultaneously. This is promising, since this experiment is a better reflection of a practical application than the other experiments, as one will rarely have all phoneme examples. It must be noted that the error values for experiment 4 have a much larger variance, since phonemes to be estimated were chosen randomly every time. The ratios between vowel and consonant occurrences are also equal, whereas the other experiments have more consonants that are estimated than vowels. The mean normalized MSE for vowels is also slightly lower than that of consonants and thus the error value in experiment 4 will tend to be slightly lower than if the conditions had been exactly the same as in the other experiments.

Although there is a slight decrease in the overall MSE when only within-class information is used for duration estimation (Exp. 2), this improvement is not uniform. A combined analysis of experiments 2 and 3 and the correlation coefficients indicates that the cross-class correlations are the reason for this behaviour. Examples include vowels such as “ae” and semivowels such as “r” and “l”.

Note that the nearest neighbor approach performs significantly worse than the other two methods. This may seem counterintuitive, but if there is limited (or even negative) correlation between the estimated and nearest neighbor estimates, one can easily see that larger error values will be observed in this case. If we let x_1 be the duration we want to estimate and x_2 the predictor, the expected value of the MSE can be expressed as

$$< (x_1 - x_2)^2 > \tag{3}$$

Multiplying out gives

$$< x_1^2 - 2x_1x_2 + x_2^2 > \tag{4}$$

Subtracting the mean value from x_1 and x_2 respectively will not change the expected value. It then follows that

$$< x_i^2 >= \mu_i^2 + \sigma_i^2 \tag{5}$$

but $\mu_i^2 = 0$ since the means are subtracted, giving $< x^2 >= \sigma^2$. (5) can be rewritten as

$$\sigma_1^2 - 2 < x_1x_2 > +\sigma_2^2 \tag{6}$$

For the global mean case this is equivalent to

$$2\sigma_1^2 - 2 < x_1x_2 >, \tag{7}$$

where we have assumed that the training speakers and testing speakers all have roughly the same variance per phoneme ($\sigma_1 \approx \sigma_2$).

From the above analysis it can be seen that for large correlations the error will tend to zero, but for small correlations the global mean approach will tend to have double the error of the nearest neighbor approach.

The variance-normalized MSE values for all phonemes in experiment 1 are shown in Fig. 5 and Fig. 6. Fig. 5 shows that

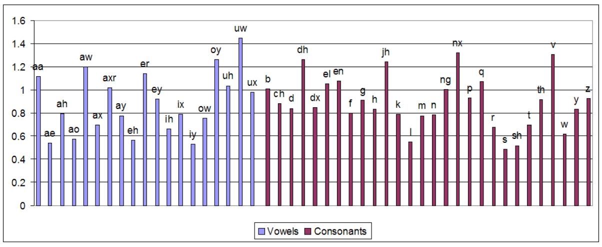

Figure 5: *Variance normalized MSE for the different phonemes using the ML approach.*

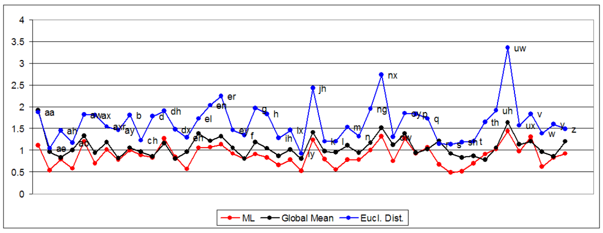

Figure 6: *Variance normalized MSE for the different phonemes using the ML, global mean and Euclidean distance approaches.*

there are significant differences in the relative predictabilities of the different phonemes, with the vowels “iy”, “eh”, “ae”, the fricatives “s”, “sh” and the liquid “l” being most predictable. These are also the phonemes whose durations correlate most strongly with those of other phonemes. The least predictable phonemes are characterized by factors such as data scarcity (“oy”), phonemic ambiguity (“uw”) and weak correlation with other phoneme durations (“nx”). It is interesting that the plosives “t”, “k”, and “d” are fairly predictable, whereas the other three plosives are less so.

The results of experiment 5 are summarized in Figures 7 to 10. As expected the error value decreases rapidly when the phonemes with the highest correlation are given as examples. An unexpected phenomenon is that even the highly predictable phonemes’ errors start to increase after a moderate number of phonemes have been added as examples. This is probably a result of the Gaussian distribution, which is assumed during our ML estimation, and deserves further attention.

6. Discussion and conclusions

The pareto chart in Fig. 1 is a confirmation of the claim that much of the variation observed in the duration of phonemes, as caused by the variable “speaker”, can be explained by a relatively small number of factors. Figures 2, 3 and 4 show that a common lengthening or shortening of all phonemes is the strongest single effect, but that differential stretches between and within phoneme classes also play a significant role.

This knowledge was then applied by estimating an ML model from the training data in the TIMIT corpus. The model was tested using the testing data, also from the TIMIT corpus. From Table 1 and Figure 6 it can be seen that the ML approach performs significantly better than the mean phone duration approach. Thus, the observed intra-speaker correlations between phoneme durations are practically usable.

High correlations between phonemes in the same class, but also across classes were observed. It was found that

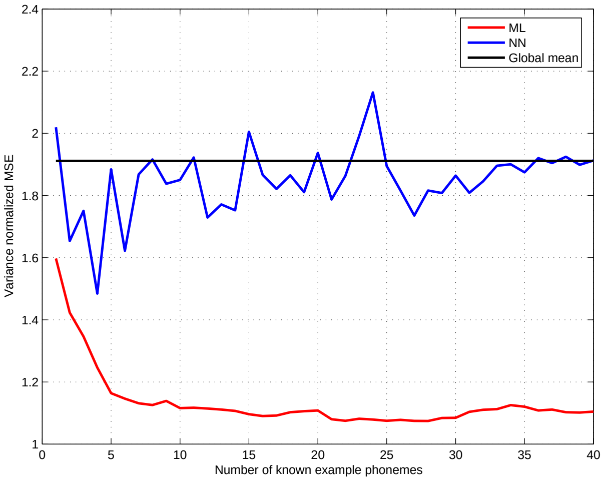

Figure 7: *Variance normalized MSE for aa vs the number of known phonemes during estimation, added in descending order of correlation with aa.*

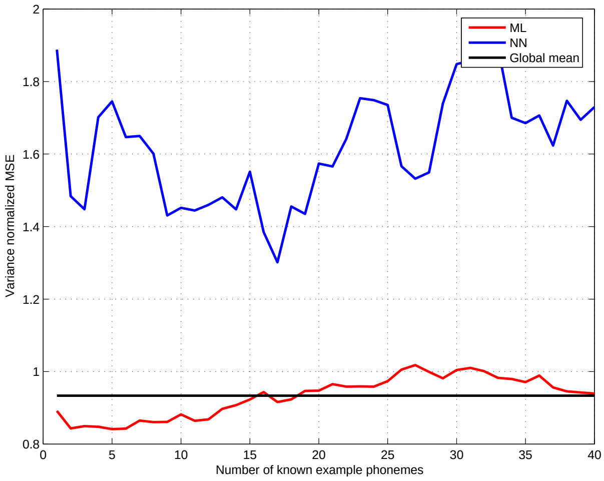

Figure 8: *Variance normalized MSE for p vs the number of known phonemes during estimation, added in descending order of correlation with p.*

most phonemes correlate well with only a few other phonemes (on the order of 10), and that accurate duration estimation is achieved using only those phonemes. As can be seen in Table 1, the lowest achievable error rate when selecting input phonemes in this fashion is 0.815, approximately 6.5% better than the result from experiment 2.

Our results also emphasize the importance of combining the various effects that influence the durations of phonemes. We found that about 15 % to 20 % of the intra-speaker variability in phoneme durations can be explained without reference to other factors, which indicates a significant role for those factors.

7. References

[1] L.C.W. Pols, X Wang, and L.F.M. ten Bosch, “Modelling of phone duration (using the TIMIT database) and its potential benefit for asr,” *Speech Communication*, pp. 161–176, 1996.

[2] C.J. van Heerden and E. Barnard, “Speech rate normalization used to improve speaker verification,” in *Proceedings of the Symposium of the Pattern Recognition Association of South Africa*, South Africa, 2006, pp. 2–7.

[3] C.J. van Heerden and E. Barnard, “Speaker classification ii, selected projects,” vol. 4441 of *Lecture Notes in Computer Science*, chapter Durations of Context-Dependent Phonemes: A New Feature in Speaker Verification, pp. 93–103. Springer, February 2007.

[4] H.R. Pfitzinger, “Intrinsic phone durations are speaker-specific,” in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, September 2002, vol. 1, pp. 1113–1116.

[5] T.H. Crystal and A.S. House, “Segmental durations in connected-speech signals: syllabic stress,” *The Journal of the Acoustical Society of America*, pp. 1574–1585, 1988.

[6] L. Ferrer, H. Bratt, V.R.R. Gadde, S. Kajarekar, E. Shriberg, K. Sönmez, A. Stolcke, and A. Venkataraman, “Modeling duration patterns for speaker recognition,” *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 2017–2020, 2003.

[7] E. Shriberg, L. Ferrer, S. Kajarekar, A. Stolcke, and A. Venkataraman, “Modeling prosodic sequences for speaker recognition,” *Speech Communication*, pp. 455–472, February 2005.

[8] V. R. Rao Gadde, “Modeling word duration for better speech recognition,” in *Proceedings of the NIST Speech Transcription Workshop*, May 2000.

[9] E. Shriberg and L. Ferrer, “A text-constrained prosodic system for speaker verification,” in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, August 2007, vol. 1, pp. 1226–1229.

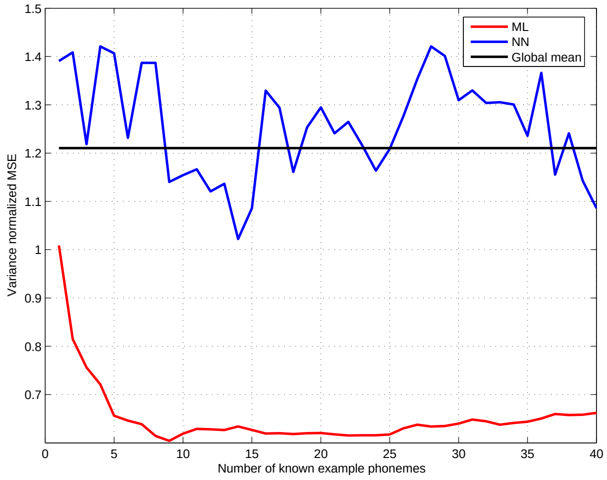

Figure 9: *Variance normalized MSE for r vs the number of known phonemes during estimation, added in descending order of correlation with r.*

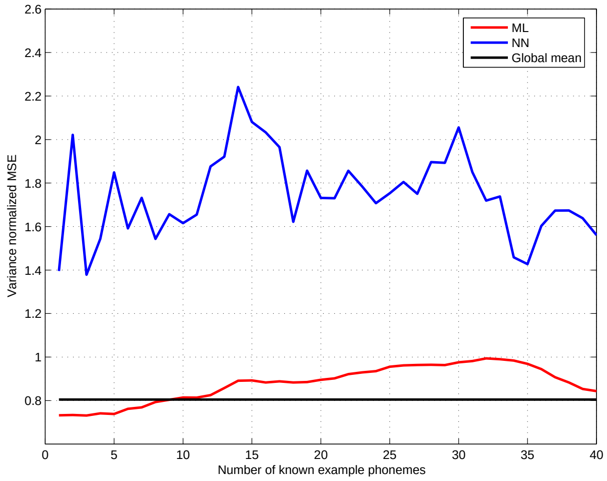

Figure 10: *Variance normalized MSE for dx vs the number of known phonemes during estimation, added in descending order of correlation with dx.*

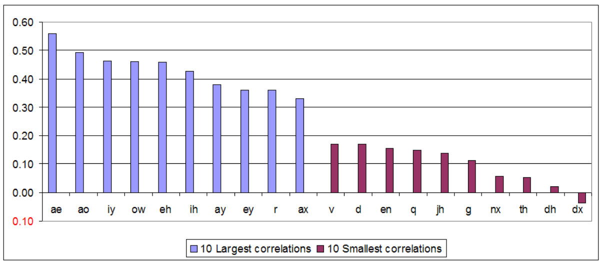

Figure 11: 10 *phonemes with the highest and* 10 *with the lowest correlation with aa.*

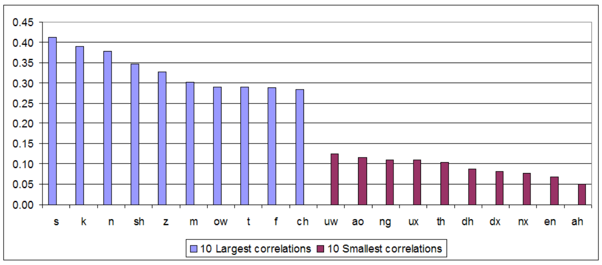

Figure 12: 10 *phonemes with the highest and* 10 *with the lowest correlation with p.*

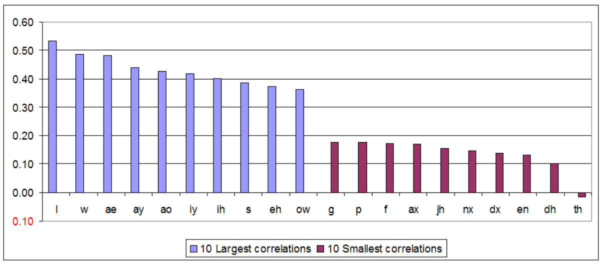

Figure 13: 10 *phonemes with the highest and* 10 *with the lowest correlation with r.*

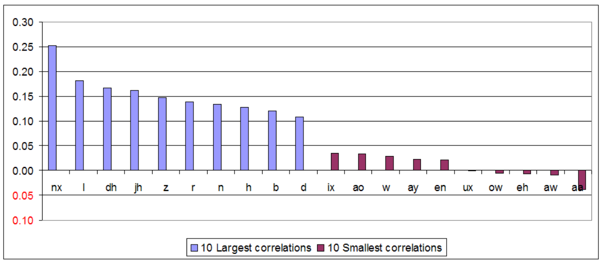

Figure 14: 10 *phonemes with the highest and* 10 *with the lowest correlation with dx.*