PREDICTABILITY OF THE EFFECTS OF PHONEME MERGING ON SPEECH RECOGNITION PERFORMANCE BY QUANTIFYING PHONEME RELATIONS

Lia Saki Bučar Shigemori, Uwe D. Reichel, Florian Schiel

Institute of Phonetics and Speech Processing, University of Munich
{lia, reichelu, schiel}@phonetik.uni-muenchen.de

Abstract: To investigate whether the impact of phoneme merging on recognition rate can be predicted, different measures to quantify the relationship between two phonemes *a* and *b* were compared: (1) the functional load of their opposition, (2) the bigram type preservation, (3) their information radius, (4) their distance within an information gain tree induced from a distinctive feature matrix, and (5) the symmetric Kullback-Leibler divergence. For each of 25 phoneme pairs we trained a speech recognizer on data in which the respective pair was merged. Based on correlation analyses and predictor selection in stepwise regression modelling we found that the impact of phoneme merging on accuracy can tentatively be captured in terms of functional load and tree distance between the merged phonemes.

1 Introduction

Large lexicon speech recognition systems use an acoustic model based on subword units, usually a phoneme set. The selection of phonemes to be modeled is crucial for the performance of the system. It should be able to represent and distinguish the sound units contrastive in a language but still allow variation in speech. Merging phonemes reduces the number of subword unit models needed to be trained and stored, solving the data sparseness problem and ideally also reducing the number of confusable phonemes, but increases the number of homophones and with it the ambiguity [1, 2]. In this paper we want to investigate whether the impact of phoneme merging on recognition rate can be predicted, comparing several kinds of quantified relationships between phonemes.

2 Data

2.1 Data for HMM training

The speech recognition system was created using the HMM toolkit (HTK) [3]. We used the Verbmobil 1 (VM1) speech corpus, separated into training, development and test corpus as suggested by BAS [4]. The phonemes were modeled with left-to-right topology HMMs, diphthongs using four states, long vowels using three and all other sounds using two states. MFCC, Energy, first and second derivation were represented by a multivariate Gaussian distribution.

2.2 Data for phoneme system quantification

For examining relations of phoneme pairs within a phonological system we used parts of the VM1 corpus comprising 285 280 word tokens with their canonical transcriptions that in total consist of 1 635 406 phoneme tokens.

3 Method

3.1 Quantification of relations in phoneme systems

We utilized five measures to represent the relation between phonemes a and b in canonical word-segmented transcriptions: (1) the functional load of their opposition, (2) the bigram type preservation after merging a and b, (3) their information radius, (4) their distance within an information gain tree induced by a distinctive feature matrix, and (5) the symmetric Kullback-Leibler Divergence.

3.1.1 Functional Load

In general the functional load FL of a phonological opposition of the phonemes a and b is related to the number of contrasts this opposition is responsible for in a language L. The basic information theoretic definition adopted here was first introduced in [5]:

$$FL(a,b) \quad = \quad \frac{H(L) - H(L_{a=b})}{H(L)}. \tag{1}$$

$H(L)$ is the entropy of a language L. $L_{a=b}$ denotes a language lacking an opposition of a and b. $FL(a,b)$ thus stands for the relative amount of information loss resulting from such a merging, reflecting the increase of homophones.
L was given by all transcription types of our data, thus by a pronunciation dictionary.

Hypothesis H1 A high functional load of the opposition of the phonemes a and b is an indicator of decreasing recognition performance, since a high degree of ambiguity is added to the data after merging a and b.

3.1.2 Bigram type preservation

By calculating the functional load on the basis of pronunciation dictionaries it is not possible to make use of disambiguating context. However, this context, which is available in connected speech, weakens the disturbing impact of homophony resulting from phoneme merging.
Thus to get a more realistic notion of the effect of merging phonemes a and b than provided by the functional load, we measured the preservation of transcription word bigram types BP as follows:

$$BP(a,b) \quad = \quad \frac{T_{a=b}}{T}, \tag{2}$$

T is the number of word bigram types in our data, and $T_{a=b} \leq T$ is the number after merging phonemes a and b.

Hypothesis H2 As for the functional load again low bigram type preservation, thus low $BP(a,b)$ values, are expected to indicate decreasing recognition performance due to increasing ambiguity.

3.1.3 Information radius

To account for the distributional distance between phonemes a and b, we used the information radius IR metrics. This measure is already established on various linguistic levels, e.g. to quantify semantic similarity [6] or the similarity of theater plays [7]. It quantifies the difference between the phoneme probability distribution p in the context of a as opposed to q in the context of b as follows:

$$IR(p,q) = D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2}), \text{ where} \quad (3)$$

$$D(p||q) = \sum_i p_i \log_2 \frac{p_i}{q_i}. \quad (4)$$

In our approach the context has been defined as the phoneme history in a phoneme bigram model. The relative entropy $D(p||q)$ gives the number of bits additionally needed to encode events i, for which the distribution p holds, by a code based on q. $IR(p,q)$ is a symmetric version of this divergence measure and thus a proper distance metrics.

Hypothesis H3 Opposed to a high functional load, a high information radius of two phonemes indicates that their merging only has a low effect on ambiguity, which qualifies them to be good merging candidates.

3.1.4 Information gain tree distance

From the distinctive feature matrix given by Chomsky [8] we induced an information gain tree classifier similar to the C4.5 algorithm by [9]. The feature matrix thus was used as training data for this tree classifier for phoneme categories, and the phonemes are represented as paths through this tree to a leaf carrying the phoneme label. The German consonant tree is shown in Figure 1. The tree representation of a phoneme set has three advantages compared to the matrix representation. First, the tree is created by recursive partitioning of the phonemes with respect to the feature which contributes the highest information gain about the phoneme identity. This allows the features to be ordered by their contribution in phoneme identification: the higher a feature's contribution, the higher its position within the tree. Second, by this splitting criterion, features that do not contribute anything to phoneme distinction, are discarded, which is helpful, if one works only on a phoneme subset. Third, due to the termination criterion, that blocks further tree splitting in case no further division of the data is possible, potential insufficiencies of a feature system are depicted as multiple phoneme labels ending at the same leaf.

Since the information gain tree was designed to represent the phoneme system exhaustively in an interpretable way in contrast to the C4.5 algorithm no pruning was carried out, and distinctive features were not allowed to be used more than once during tree construction.

The tree distance TD between two phonemes finally was defined as the number of edges to be passed on the shortest path from one phoneme to the other.

Hypothesis H4 Since the acoustic similarity of two phonemes is to some extent reflected by their closeness in the tree, a low distance of two phonemes should indicate a relatively good compatibility of the related acoustic models for speech recognition, which would qualify them to be merged.

```
ant=0
|  back=0
|  |  cont=0
|  |  |  grave=0: C
|  |  |  grave=1: Q
|  |  cont=1
|  |  |  cons=0: j
|  |  |  cons=1
|  |  |  |  cor=0: h
|  |  |  |  cor=1: S
|  back=1
|  |  cont=0
|  |  |  voi=0: k
|  |  |  voi=1: g
|  |  cont=1
|  |  |  delrel=0: R
|  |  |  delrel=1
|  |  |  |  son=0: x
|  |  |  |  son=1: N
ant=1
|  cont=0
|  |  cor=0
|  |  |  nas=0
|  |  |  |  voi=0: p
|  |  |  |  voi=1: b
|  |  |  nas=1: m
|  |  cor=1
|  |  |  nas=0
|  |  |  |  voi=0: t
|  |  |  |  voi=1: d
|  |  |  nas=1: n
|  cont=1
|  |  cor=0
|  |  |  voi=0: f
|  |  |  voi=1: v
|  |  cor=1
|  |  |  delrel=0: l
|  |  |  delrel=1
|  |  |  |  voi=0: s
|  |  |  |  voi=1: z
```

Figure 1 - Compact information gain tree representation of the German consonant inventory in terms of the distinctive features **ant***erior, back,* **cont***inuous, grave,* **cons***onantal,* **cor***onal,* **voi***ced,* **del***ayed* **rel***ease,* **son***orant* [8]. To be read from left to right, e.g. */g/* is *back* and *voiced* (row 14). The within-tree distance TD, i.e. the path length, between */l/* and */v/* amounts 4 (2 edges up, and 2 edges down). German SAMPA is used for the phoneme transcriptions.

3.1.5 Symmetric Kullback-Leibler Divergence

To measure the similarity between phonemes a and b on signal level, the symmetric Kullback-Leibler divergence (KL) [10] was used:

$$D_{KL}(a,b) = \frac{1}{2}tr[(\sum\nolimits_a - \sum\nolimits_b)(\sum\nolimits_b^{-1} - \sum\nolimits_a^{-1})] + \frac{1}{2}tr[(\sum\nolimits_a^{-1} + \sum\nolimits_b^{-1})\delta\delta^T] \quad (5)$$

$\sum_a$ and $\sum_b$ are covariance matrices of the phonemes a and b, δ is the difference in the means and tr the matrix trace function. In short, this measure compares the form and size of two probability density functions.

In the speech recognition system we used for the experiments, diagonal variance vectors were used, so the KL was calculated on another HMM model, trained on the same data as the baseline system but where the states were defined using covariance matrices.

Hypothesis H5 A smaller *KL* indicates greater similarity between two models, and thus a greater confusability. Assuming that two phonemes with a smaller *KL* are more often mismatched, merging them should have a smaller impact on speech recognition rate.

To summaries, we expect that phonemes *a* and *b* with

1. low functional load,
2. high bigram type preservation,
3. high information radius,
4. low information tree distance, and
5. low symmetric Kullback-Leibler divergence

are good candidates for merging, since their acoustic models are expected to be close, and their merging should increase ambiguity only to a small extent.

3.2 Speech recognizer training and evaluation

The baseline system included 47 phonemes, of which 5 represent non-speech voices as laughter, breathing or background noise, as listed in table 1. The accuracy rate of the baseline system was 69.44. A pair of consonants or vowels was picked to merge and the new speech recognition system was trained with 46 phonemes. We did not merge a consonant with a vowel.

Vowels	e, 2:, y:, OY, E:, 9, Y, aU, o:, u:, O, U, e:, aI, E, i:, @, a:, a, I, 6
Consonants	p, S, N, j, h, x, g, k, b, z, r, C, ?, l, f, v, d, m, s, t, n
Other sounds	<usb>, <nib>, <la:>, <br:>, <p:>

Table 1 - Phonemes used in the baseline system.

4 Results

We retrained the baseline recognition system 25 times, each time with a different phoneme pair merged and compared the resulting accuracy rates with the calculated measures. The results are shown in Table 2.
The correlations between recognition accuracy and the phoneme relations are visualized in Figure 2.

1. Functional load showed negative correlation with recognition accuracy with p=0.12, tentatively confirming hypothesis $H1$.
2. Bigram type preservation correlated positively with recognition accuracy, although with p=0.18 only tentatively confirming hypothesis $H2$.
3. Information radius and recognition accuracy showed positive correlation with p=0.59, tentatively confirming hypothesis $H3$.

pairs	accuracy	KL	IR	FL	BP	TD
2: 9	69.47	1883410.79	1.0783860	0	1.0000	2
Y y:	69.42	1363217.07	0.7483130	5.736440e-16	1.0000	2
u: U	69.35	1693574.93	1.3002731	1.099554e-04	1.0000	2
E: E	69.29	1784305.47	1.5056551	0	0.9998	2
Y I	69.24	1333615.65	0.9648393	0	0.9999	4
y: i:	69.23	4473329.68	0.6242079	3.785046e-04	0.9994	4
a: a	69.20	1503416.63	1.2532038	1.477526e-03	0.9993	2
o: U	69.17	1777095.58	1.0230573	0	1.0000	5
N n	69.17	1435198.65	0.6401513	3.615079e-04	1.0000	9
k g	69.04	78259.46	1.1946101	1.371553e-03	0.9997	2
z s	68.97	179217.27	1.4006399	4.030575e-04	0.9999	2
p k	68.96	104196.39	0.5859276	8.278275e-05	1.0000	9
I e:	68.95	1575669.53	1.0781021	8.532797e-04	0.9998	6
r l	68.83	3053839.46	1.0000000	0	1.0000	8
h x	68.81	1957869.21	1.9596388	0	1.0000	8
k h	68.74	410403.89	1.4918656	1.729387e-04	0.9998	7
b v	68.59	272070.26	1.2759282	1.388183e-03	0.9994	7
f h	68.56	1106008.13	1.1170951	1.577609e-03	0.9994	9
m n	68.52	1232643.01	0.5891027	1.041849e-02	0.9842	4
l v	68.38	2813286.06	0.9876505	1.014802e-03	0.9994	4
v f	68.33	913065.62	0.7554164	2.305747e-03	0.9981	2
d t	68.28	28734.19	1.1907792	5.849359e-04	0.9997	2
s S	68.26	1262620.93	1.0084304	6.027621e-04	0.9997	10
l k	68.21	2748864.04	0.7052246	2.226014e-04	0.9999	8
s f	67.95	714748.74	0.9178399	1.630141e-03	0.9970	5

Table 2 - Recognizer accuracy and phoneme relations. *KL*: Kullback-Leibler divergence, *IR*: information radius, FL: functional load, *BP*: bigram type preservation, TD: information gain tree distance.

4. Hypothesis $H4$ about lower information tree distance indicating smaller declination of accuracy was confirmed by our data showing negative correlation (-0.3450) with $p \leq 0.1$.

5. Hypothesis $H5$ was not confirmed, as the Kullback-Leibler divergence and recognition accuracy showed positive correlation (0.1895) with p=0.36.

To quantify the strength and the direction of the influence of a phoneme relation on recognition accuracy we applied a linear regression on the z-normalized data in table 2 using the phoneme relations as predictors of accuracy. The regression coefficients shown in table 3 contradict the hypotheses $H2$, $H3$, and $H5$ related to *BP*, *KL*, and *IR*, respectively, which made us to expect opposite algebraic signs. FL and TD in contrast show an influence in the expected direction confirming hypotheses $H1$ and $H4$.

predictor	FL	BP	IR	TD	KL
weight	-1.1692	-0.7610	-0.0610	-0.3896	0.1193

Table 3 - Linear regression coefficients for the z-normalized phoneme relations indicating their influence direction and strength on recognition accuracy.

In accordance with the regression coefficient in a stepwise linear regression only FL and TD were selected as predictors providing additional confirmation of hypotheses $H1$ and $H4$.

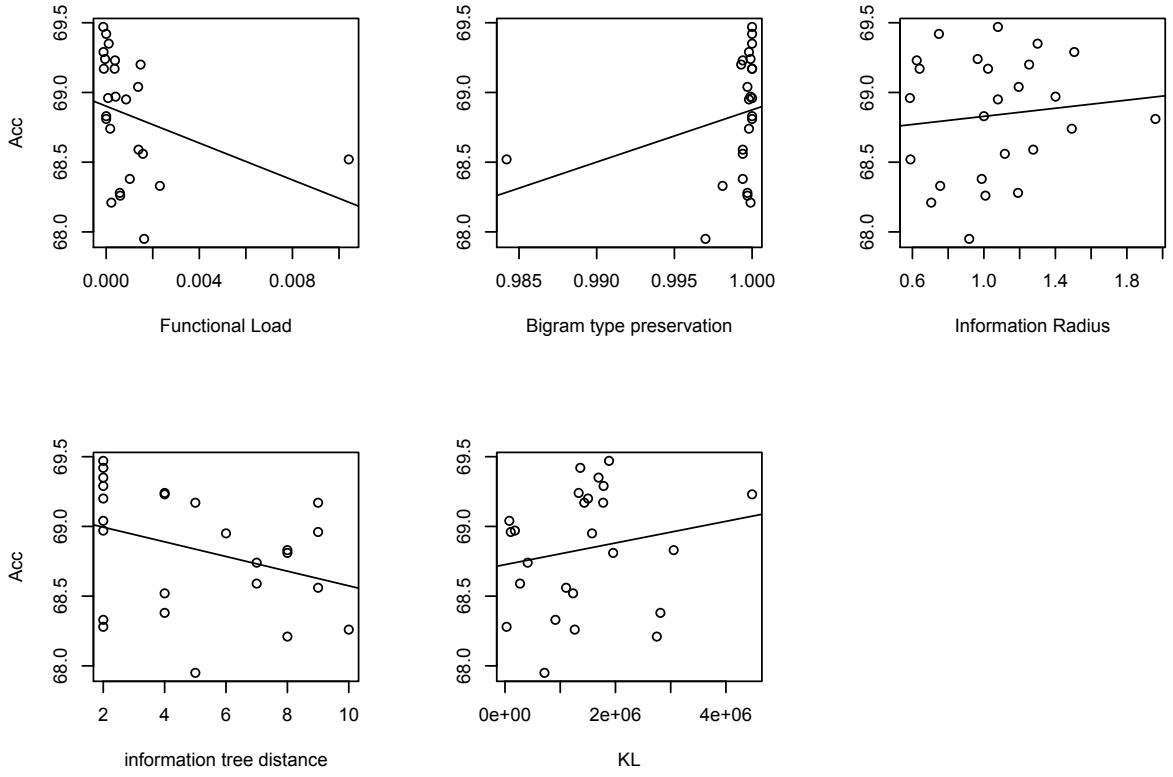

Figure 2 - Illustration of dependency between the Accuracy and the phoneme distances.

In summary, as shown in table 4 only FL and TD turned out to capture the impact of phoneme relations on recognition performance.

hypothesis (measure)	correlation	regression weight	predictor selection
$H1$ (FL)	$+$	$+$	$+$
$H2$ (BP)	$+$	$-$	$-$
$H3$ (IR)	$+$	$-$	$-$
$H4$ (TD)	$+$	$+$	$+$
$H5$ (KL)	$-$	$-$	$-$

Table 4 - Confirmation overview of the measure-related hypotheses formulated in section 3.1.

5 Discussion

Using five different measures to quantify phoneme relations, we tried to find out if it is possible to predict the impact of merging a phoneme pair on the accuracy of a speech recognition system. The functional load of a phonological opposition of phonemes a and b is related to the number of contrasts that this opposition is responsible for in a language. The bigram type preservation is similar to the functional load but also takes the word context into account. The information radius accounts for the distributional differences of a and b with respect to their phoneme contexts. These three measures are based on lexical data. The information gain tree is based on the phonological knowledge of the language, and on a distinctive features matrix. The symmetric Kullback-Leibler divergence measures the similarity between phonemes on the signal level.

Only the functional load and the information gain tree distance turned out to have an impact on accuracy after phoneme merging. The low performance of the Kullback-Leibler divergence could be attributed to the fact that it has been calculated on a different model. Testing the impact of merging on the model *KL* was calculated on could answer this question. Nevertheless, in case *KL* is a better accuracy predictor in the context of another model, this indicates, that for different HMM state models, different phoneme sets should be used. Since the information gain tree can to some degree predict the impact of merging phonemes, we can conclude that a deeper understanding of the phoneme system of a language is of relevance to set up a speech recognition system.

References

[1] D. Vazhenina and K. Markov, "Phoneme set selection for russian speech recognition," in *Natural Language Processing andKnowledge Engineering (NLP-KE), 2011 7th International Conference on*, nov. 2011, pp. 475 –478.

[2] J. song Zhang, X.-H. Hu, and S. Nakamura, "Automatic derivation of a phoneme set with tone information for chinese speech recognition based on mutual information criterion," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, may 2006, p. I.

[3] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[4] "http://www.phonetik.uni-muenchen.de/bas/basvm1eng.html," Verbmobil I web page.

[5] C. Hocket, "The quantification of functional load," *Word*, vol. 23, pp. 320–339, 1967.

[6] C. Manning and H. Schütze, *Foundations of statistical natural language processing*. Cambridge, Massachusetts: MIT, 2001.

[7] U. Reichel, "Informationstheorie in der Literaturwissenschaft. Ein Beitrag zur Shakespeareforschung," in *CLARIN-D Newsletter*, 2012, vol. 2, pp. 19–20.

[8] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper & Row, 1968.

[9] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993.

[10] J. Campbell, J.P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437 –1462, sep 1997.