Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 24 (2001), 75-84.

ALTERNATIVES IN TRAINING ACOUSTIC MODELS FOR THE AUTOMATIC RECOGNITION OF SPOKEN CITY NAMES

Daniel S. Salomons & Louis C.W. Pols

Abstract

Training the acoustic models for automatic speech recognition (ASR) as well as the similarity between the training corpus and the recognition task have a major influence on the performance of a speech recogniser. The more similar the two sets are, the better the performance of the speech recogniser will be. When the recognition task consists of city names, this implies that the training corpus must consist of city names only. This causes problems, especially in the Netherlands, because there are not enough speech data available of the smaller cities or villages to satisfy the need for rare phonemes. The usual alternative in such a case is training acoustic models with a speech corpus consisting of phonetically rich sentences. The disadvantage is that these sentences are spoken in a 'read aloud' speech style, while the intended recognition task consists of spontaneous speech. This causes a great decrease in performance. Adding (sur)names, street names and application words only adds a small number of rare phonemes to the training data. Fewer speech data are generally a performance-decreasing factor in speech recognition. On the other hand the greater similarity with the recognition task is a performance-increasing factor. In this research the subject of investigation was whether this latter factor would outweigh the former one in this specific task. This appeared to be the case. This research was done during an intern at KPN Research in Leidschendam, the Netherlands. This intern was part of the master's graduation project of the first author of this paper. The report about this project was also his master thesis (Salomons, 2000).

1 Introduction

A phone directory service is very labour intensive, which makes it a very expensive service. Therefore, phone companies have been investigating the capabilities of ASR to automate such services. In the beginning of 1999 the customer interaction of the national phone directory service was automated. This implied that from then on the data needed for finding the required phone number, as well as the feedback to the customer, were done automatically. The actual transcription of these data was still done by operators, listening to the customer's speech, and entering the data. Experiments to automate also this task, by applying ASR to recorded speech samples, showed that this customer interaction design was not yet suitable with current ASR technology. Therefore a different customer interaction procedure was designed. The required data, surname, and eventually street name and city name, were to be asked and recorded separately. This approach seemed to be a lot more hopeful than the

previous one. Especially applying ASR for city names seemed to be very promising. Also the set of city names in the Netherlands is rather constant and limited, between 2,350 and 2,400. This was the starting point of the current project.

When training acoustic models for a specific recognition task it is important that the training material is similar to the kind of speech that is going to be used in the actual recognition task. The greater the similarity, the better the speech recogniser will perform. This means that for a task consisting of spontaneously spoken city names, the training data must contain (only) city names as well. The problem is that there is currently not a sufficient amount of such speech data available to put together a training corpus that will let the speech recogniser perform satisfactory. In such cases usually a training corpus consisting of phonetically rich sentences is used to train the acoustic models. In phonetically rich sentences an effort has been made to get a sufficient amount of all phonemes (also the rare ones) in the speech database, which is a necessary condition to train models for every phoneme. However, such sentences have some major disadvantages as well. Most important of all, they are read from paper, which makes the speech style quite different from the speech style in the intended recognition task.

At KPN Research the speech database with names of the current dwelling place was used as a test case to automatically recognise city names, because their speech style is very similar to the recognition task in the phone directory service mentioned above. However, training acoustic phone models then caused a dilemma. Using only the city names from the public part of the available database did not supply (enough) data for training all phonemes, whereas using phonetically rich sentences decreases the speech recogniser's performance due to dissimilarity in speech style. At some moment the idea rose that adding application words from the public part of the speech database, plus surnames and street names from the confidential part of the speech database might supply enough occurrences of rare phonemes to train models for them. Also the greater spontaneity of this material would increase the similarity regarding speech style with the recognition task. On the other hand this would still provide far less training data than in the corpus of phonetically rich sentences. The question was which factor would outweigh the other in this particular case, similarity in speech style or amount of training data for rare phonemes.

1.1 Speech Recogniser

The speech recogniser used in this research was Phicos, a Philips recogniser for research purposes. It can be configured in many ways. In this research it was configured as a serial tied state recogniser with monophone models with a limited skip and 6 states. More details about the Philips speech recognition technology have been published by Steinbiss et al. (1995).

2 Training Material

For all experiments described here the training corpora were selections from the Dutch Polyphone. A short description of this database will be given below, followed by a description of the actual selections.

2.1 Dutch Polyphone

In the 1990's KPN Research and SPEX recorded the Dutch Polyphone corpus. Participants were asked to answer over the telephone some questions as printed on paper, among them the question about the city where they were born and where they had lived. They also read some sentences, among them five phonetically rich sentences and some application words. All these speech data were included in the release of the Dutch Polyphone, and will be called the public part of Polyphone. Utterances of 5,050 individuals are contained in this part. Table 1 provides information about the number of speakers per Dutch province.

Table 1 Number of representatives from every province in the Netherlands. Please note that amounts are on the left of each pair of columns.

number	province	number	province	number	province
260	Drente	351	Groningen	329	Overijssel
168	Flevoland	309	Limburg	421	Utrecht
298	Friesland	550	Noord-Brabant	267	Zeeland
593	Gelderland	660	Noord-Holland	843	Zuid-Holland

All participants should have spoken 44 items. 4,810 participants actually did, 236 participants missed one item, three missed two items, and of one participant only 16 items have been recorded. Of most items there are either 5,049 or 5,050 occurrences, as can be seen in Table 2.

Table 2 Number of occurrences of every item. APPSNT means application sentence, APPWRD means application word, GLDAMT means guilder amount, PHONSNT means phonetically rich sentence, SPWORD means spelled word, YN means yes/no answer.

number	item	number	item	number	item
5049	AGE.WAV	5049	DATE.WAV	5049	PHONSNT4.WAV
5046	AMOUNT1.WAV	5049	DIGITS.WAV	5049	PHONSNT5.WAV
5046	AMOUNT2.WAV	5050	EDUCAT.WAV	5049	POSTC.WAV
5048	AMOUNT3.WAV	5050	GENDER.WAV	4869	REMARK.WAV
5049	APPSNT1.WAV	5046	GLDAMT1.WAV	5047	SPWORD1.WAV
5049	APPSNT2.WAV	5049	GLDAMT2.WAV	5049	SPWORD2.WAV
5049	APPSNT3.WAV	5048	GLDAMT3.WAV	5049	SPWORD3.WAV
5049	APPSNT4.WAV	5048	NUMBER1.WAV	5044	TIME1.WAV
5050	APPWRD1.WAV	5047	NUMBER2.WAV	5026	TIME2.WAV
5049	APPWRD2.WAV	5044	NUMBER3.WAV	5050	YN1.WAV
5049	APPWRD3.WAV	5049	NUMBER4.WAV	5048	YN2.WAV
5049	APPWRD4.WAV	5045	NUMBER5.WAV	5049	YN3.WAV
5050	CITY1.WAV	5049	PHONSNT1.WAV	5050	YN4.WAV
5050	CITY2.WAV	5049	PHONSNT2.WAV	5050	YN5.WAV
5049	CITY3.WAV	5049	PHONSNT3.WAV		

These and other counts have been achieved by using UNIX commands, combined with pipelines.

Participants were also asked to mention their surname, street name, house number, postal code and dwelling place. Due to privacy legislation the latter speech data are strictly confidential, and cannot be used outside KPN Research or SPEX. There are 5,375 occurrences of the surname item, of which 277 are transcribed as `*missing*`, which means that there are no speech data in the corresponding speech data file. There are 5,374 occurrences of the street name item, of which 277 are also transcribed as `*missing*`. These speech data will be called the confidential part of Polyphone. The speakers in both parts are mainly the same group, but the connection between the

speech data of a participant in either part could not be re-established. A more extensive description of the public part of the Dutch Polyphone can be found in Den Os et al. (1995).

2.2 Phonetically Rich sentences

Each of the participants in the Polyphone corpus read aloud five phonetically rich sentences. From the training corpus all utterances with background noise, background speech, clicks, uh's and utterances that could not be understood by the transcriber, were excluded. This resulted in a training corpus of 23,499 utterances that were used to train the acoustic phone models. The phoneme distribution can be found in Table 6 in the Appendix.

2.3 Short Utterances

The surname and street name items from the confidential part of Polyphone, and the application words and city names from the public part, are mainly utterances with not more than four words. Only 2.9 % of the utterances contain more than four words. Therefore we call this the short utterances training corpus.

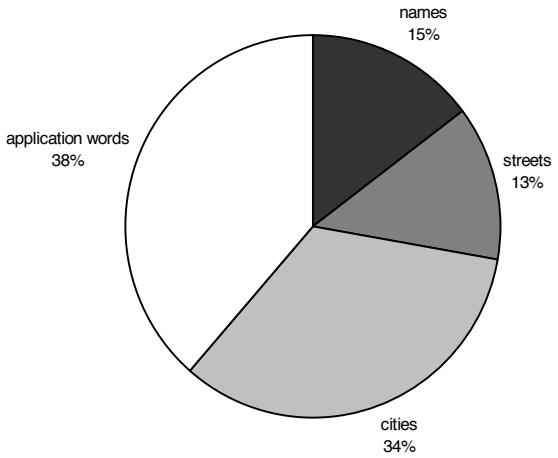

Figure 1 Ratios of items in the short utterance training corpus

The available pronunciation lexicon provides a phonetical transcription in SAMPA format, except that postvocal /l/ and postvocal /r/ were denoted by us as /L/ and /R/, respectively. Two phoneme models were trained for both phonemes, a prevocal one and a postvocal one. The lexicon was derived from Celex and was included in the Polyphone release, as far as the items from the public part are concerned. For the items from the confidential part (surnames and street names), the lexicon was

completed with information from the Onomastica project. For a few hundred words there was no entry in either source lexicon, so the first author made a phonetical transcription by hand. Thereupon postvocal /l/ and postvocal /r/ symbols were changed to /L/ and /R/ respectively, which is not standard SAMPA.

3 Test

The test corpus and the test conditions as described below were originally designed in August 1997 at the 'Instituut voor Taal, Spraak en Informatica' (Institute for Language, Speech and Computer Science' of the 'Katholieke Universiteit Nijmegen' (Catholic University Nijmegen). The test was used for many experiments, also at KPN Research. It has been used by us as well for comparability reasons.

3.1 Corpus

To make comparison with previous tests done at KPN Research, the test conditions had to be exactly the same. Therefore we did not modify the test corpus. The description of the test corpus is based on an internal report of KPN Research, which is not publicly available. The test corpus contains the city items from the confidential part of Polyphone. It was the most similar material available at the time when our experiments to automatise the city recognition task in the phone directory service were performed. Each item contains one city name only. The only aspect different from the ultimate recognition task is that the item contains the name of the city in which the participant is living, while in the real recognition task this is likely not the case. Then it will generally be a city not included in the phone directory book distributed among subscribers, which contains only phone numbers of the area in which it is distributed, and they need the phone directory service for cities not included in that book. The test corpus contained only utterances with one complete city name, plus eventually some additional noise. Utterances with a truncated city name, such as …otterdam (Rotterdam) or …ouda (Gouda), were removed from the selection. All city names are considered to be one word, and white spaces have been replaced by underscores '_'. For instance Alphen aan den Rijn is transcribed as alphen_aan_den_rijn. Sometimes participants said not just the place they are living in, this is one reason why some utterances contain more than one word. The words that were not city names, were indicated not to be counted as wrong if not recognised correctly. In Table 3, in which information about the number of words per utterance is given, these utterances were counted as empty.

The total number of city name items in the confidential part of Polyphone is 5,098. Five utterances contain no speech data, the transcription then states `*missing*`. So the total number of utterances containing speech data is 5,093. Because in a real speech recognition application real utterances can easily be distinguished from items where something went wrong, this last number is considered to be 100 %. Of these 5,093 utterances 159 contain only additional noise and 150 contain truncated city names, so 4,784 utterances remain in the final test set, which is 93.933 % of 5,093. Despite the fact that all utterances containing only noise should have been removed, the test set still contained two utterances with only noise. They have been marked as not relevant, and were not counted as wrong. The percentile performance data must be scaled with this percentage to get a realistic impression about what performance under realistic circumstances would be possible.

Table 3 Number of words per utterance in the test corpus

words per utterance	test corpus
0	2
1	4404
2	292
3	41
4	45
total number of utterances:	4784

3.2 Lexicon

The speech recogniser cannot recognise words that have no entry in the lexicon. Thus in order to get a realistic score, all city names in the Netherlands had to be included in the lexicon. The spelling of the city names was based on the postal code directory of the Netherlands, because the language model, described hereafter, was based on it. Places with the same (sounding) name, but located in different provinces were replaced by one entry. For example elst gld (Gelderland) and elst utr (Utrecht) were replaced by elst, hengelo ov (Overijssel) and hengelo gld by hengelo, and beek gld and beek lim (Limburg) by beek. Some names were added because they were current in popular speech, like west_terschelling for terschelling_west, den_haag for 's-Gravenhage (the Hague) and den_bosch for 's-Hertogenbosch (in French: Duc le Bois). The spelling of prefixes has been adapted to Onomastica, for example s<white space> became s_, t<white space> became t_, and st<white space> became st_. Three names were added that appear in composed names in the postal code directory. Two names appeared to be transcription errors that had not been identified as such when the lexicon was compiled. These names were Houwlerwijk and Wieten, which should have been Haulerwijk and Whijtmen, respectively. The latter is locally pronounced as /wit@m/. The phonetical transcriptions were derived from Onomastica where possible. The lexicon contained 2,374 entries, including four entries for noises.

3.3 Language Model

When a simple language model to this task was compiled in August 1997 there was a problem. At that particular time there were no data available about how often each city name was asked for. It is obvious that requests for large cities will occur more often than for small villages, but there were no quantitative data. The language model was simply based on the (digital) postal code directory, because it seemed the only practical way to estimate the statistical chance for a city name. Because large cities have more streets, so more street names, and the digital postal code directory contains these names in digital form, the language model was based on it in the following way. The number of street names in a place divided by the total number of streets in the postal code directory gave a number that was taken as the statistical chance that a city name would occur.

4 Results

The results of our recognition tests can be found in Table 4 and Table 5. In both tests the language model is the same. In these tables the row headers contain the items wordgraph and best sentence. The column headers contain the items, sentence error rate (SER), word error rate (WER), substitutions (Sub), insertions (Ins), and deletions (Del). First it must be noticed that the numbers are error scores and rates, which means the lower the number, the better the performance. The wordgraph score gives the number of cases in the test set where the correct recognition was not a path in the network of possible recognitions. The best sentence score gives the number of cases where another recognition than the correct one was produced as the most likely recognition, and that most likely one would normally be regarded as the result. SER gives the number of cases where the entire 'sentence' was recognised correctly. WER gives the number of words that were not correctly recognised, and it is the summation of the substitutions, insertions and deletions. In this case the SER and WER should have been the same. This is not the case, because the simple language model allowed more than one city name to be recognised in one test utterance. It is easy to see that the speech recogniser performs much better for this specific recognition task with acoustic models trained with 'short utterances'.

Table 4 Results of the test with acoustic models trained with phonetically rich sentences

	SER	WER	Sub	Ins	Del
wordgraph	360 (7.53 %)	406 (8.49 %)	326	68	12
best sentence	1028 (21.49%)	1088 (22.75)	998	76	14

Table 5 Results of the test with acoustic models trained with 'short utterances'

	SER	WER	Sub	Ins	Del
wordgraph	192 (4.01 %)	228 (4.77 %)	167	61	0
best sentence	597 (12.48 %)	646 (13,51 %)	575	70	0

5 Conclusion

From the results it is easy to conclude that the short utterance-training corpus is a good alternative for the phonetically rich sentence-training corpus. It seems plausible to state that surnames, street names and application words are a good addition if there is an insufficient amount of training data to determine the acoustic models for the recognition of utterances with one spontaneously spoken city name.

6 Discussion

The interesting question in this context is why the acoustic models trained with 'short utterances' perform so much better than the models trained with phonetically rich

sentences, despite the fact that the latter provide far more phoneme realisations than the first, see Table 6. Two possibilities seemed plausible. The speech style is so much more similar to the speech style in the recognition task, that it outweighs the smaller amount of training data. The other possibility is that during the training stage the alignment of the phonemes fails more often. Because sentences are longer, phonemes may be aligned incorrectly. If this alignment goes wrong, the subsequent phonemes are also likely to be aligned incorrectly. The answer to this question can be found in the results of experiments that have been done at KPN Research in succession to this research. These results were presented by Sturm et al. (2000). In this research a recognition experiment was done with a test set of phonetically rich sentences. In this experiment the models trained with phonetically rich sentences outperformed the short utterance models. It is obvious that speech style is the major factor in this context.

If surnames, street names and application words are a good alternative to phonetically rich sentences to train acoustic models for the recognition of city names, it might be possible that the same is true, for the recognition of surnames or street names. If this is the case, this is very interesting, because training speaker-independent acoustic models is very time consuming, because a lot of speech data must be processed. It would mean that the same acoustic models can be used for the three items in a phone number directory service (surname, street name and city name), and thus need to be trained only once. Only the language model needs to be different. This suggestion needs empirical confirmation.

References

Os, E.A. den, T. Boogaart, L. Boves & E. Klabbers (1995) "The Dutch Polyphone corpus", *Proc. Eurospeech-95*, Madrid, vol. 1, 825-828.

Salomons, D.S. (2000). *Optimale woonplaatsherkenning voor 118; Het selecteren van een trainingscorpus en de effecten daarvan* (Rapport 32315). Leidschendam, The Netherlands: KPN Research.

Steinbiss V., Ney H., Aubert X., Besling S., Dugast C., Essen U., Geller D., Haeb-Umbach R., Kneser R., Meiere H.-G., Oerder M. & Tran B.-H. (1995) "The Philips research system for continuous-speech recognition", *Philips Journal of Research* **49**: 317-352.

Sturm, J., J.H.G. Kamperman, L. Boves & E.A. den Os (2000) "Impact of speaking Style and speaking Task on acoustic Models", *Proceedings ICSLP2000*, Beijing, China, Vol 1:361-364. Also: http://lands.let.kun.nl/literature/sturm.2000.2.pdf

Appendix

Table 6 Phoneme histogram: exact amounts. Phonetic symbols are according SAMPA, extended with /L/ and /R/ for post vocal /l/ and /r/ respectively. Symbols with = denote noise markers.

SAMPA symbol	phonetically rich sentences	short utterances
2:	5229	507
9y	6950	1752
@	138690	31017
A	37782	12836
Au	5730	1322
E	28763	11601
Ei	18305	4709
I	29922	8106
L	17601	6520
N	10001	4288
O	24526	7184
R	46464	16026
S	2398	818
Y	7563	3618
a:	28992	12325
b	18528	7000
d	54725	12356
e:	27094	8138
f	10644	3363
h	15013	5118
i	22727	6177
j	9731	2509
k	34427	12160
l	25588	10077
m	27551	11132
m=	3328	2871
n	80908	23962
n=	0	0
o:	22773	7871
p	19378	6422
r	29844	13440
s	49777	20206
s=	0	0
t	97484	26078
u	8914	2492
u=	0	0
v	29827	7586
w	20463	5780
x	40335	12073
y	6261	1255
z	18275	3418
Total	1082511	334113