Impact of speaking style and speaking task on acoustic models

Janienke Sturm[2], Hans Kamperman[1], Lou Boves[1,2], Els den Os[1]

[1]KPN Research, KPN Royal Dutch Telecom, Multi-Media Department,
St. Paulusstraat 4, Leidschendam, The Netherlands
[2]Nijmegen University, The Netherlands
Email: {J.H.G.Kamperman, E.A.denOs}@kpn.com, {Janienke.Sturm, L.Boves}@let.kun.nl

ABSTRACT

The loss in performance caused by mismatch between train and test material suggests a need for task specific acoustic models, especially for highly demanding tasks. However, since the training of these models is extremely expensive, general purpose models are more attractive. In this paper we address the impact of mismatch in speaking style and task. We trained three sets of acoustic models on data from different tasks, involving both read and extemporaneous speech. The average utterance length in the training corpora varied between 10.5 and 1.2 words. The models were tested on matched as well on very different tasks. The results suggest that general purpose models trained from short utterances are to be preferred in most spoken dialog systems. However, these models might not perform adequately in dictation tasks.

1. INTRODUCTION

Up to now all operational Spoken Dialog Systems (SDS) have been tailor made. At the level of the functional specification and the design of the dialog this will probably never change. However, today the need to tailor a SDS extends downward, perhaps even to the acoustic models. Adaptation of acoustic models is extremely expensive. This has raised interest in automatic adaptation of acoustic models, using data obtained during the actual operation of the service [1]. However, adaptation is more effective as the starting point is closer to the goal.

Recently a large number of telephone speech corpora have become available to support the training of general purpose acoustic models. The oldest examples are Macrophone (American English) and Dutch Polyphone [2]. In addition SpeechDat corpora are now available for many European languages. One of the most outstanding characteristics of these corpora is the set of phonetically rich sentences, that are specifically meant to train general purpose acoustic models. In Dutch Polyphone the average sentence length was 10.5 words. This has at least two disadvantages: the speakers may have had trouble in reading the sentences completely fluently and naturally; even if the sentences are read completely correctly, the training software may still have difficulty in finding the 'true' alignment between the signal and the transcription. This raises the question whether the phonetically rich sentences in Polyphone are indeed the best possible basis to train general purpose acoustic models.

Analysis of recorded interactions between customers and SDSs has shown that the average utterance length in most of these systems is 3.5 words (or even less). Thus, the mismatch in utterance length may affect recognition performance. This was one of the reasons why the more recent SpeechDat specifications include phonetically rich words. However, to the best of our knowledge, no results of experiments comparing recognition results with acoustic models trained on phonetically rich sentences or on phonetically rich words have as yet been published.

Most of the utterances in Macrophone, Polyphone and SpeechDat are read. However, SDS applications over the telephone typically involve extemporaneous speech, which is different from read speech in many respects, probably including the way speech sounds are pronounced. In this paper we address an application that is almost by nature characterised by very short utterances, which are seldom read, viz. automatic Directory Assistance (DA). In each dialog turn the system prompts for a single information item, for instance the name of the city, or the name of the person or organisation. The responses to these prompts consist typically of a single word, or a short phrase, and most of the time the customer will know the answers to the questions by heart. Thus, in DA the phonetics of the short and extemporaneous responses may be very different from the long read utterances on which general purpose acoustic models would be trained.

In this paper we investigate the impact of the average length of the training utterances on recognition performance for several different tasks. To that end we compare three sets of acoustic models, context-dependent models trained on the phonetically rich sentences in Polyphone, context-dependent models trained on a set of short utterances, also taken from the Polyphone corpus, and context-dependent models trained on a corpus of responses to the 'For what city?' prompt in a DA application.

The performance of the three sets of acoustic models is tested on four test sets. Two of the test sets comprise only city names. The third test set comprises phonetically rich sentences from the test part of the Polyphone corpus. The fourth corpus comprises responses to the first prompt ('From which station to which station do you want to travel?') in an operational train timetable information system.

Section 2 gives a description of the speech data used for training and testing. Section 3 describes the details of the experiments. In Section 4 the results are discussed, and Section 5 summarises the most important conclusions.

2. SPEECH DATA

2.1. Training on phonetically rich sentences

The phonetically rich sentences selected for training are taken from the Polyphone train set. Each of the 4,051 speakers read five sentences, for a total of 20,110 utterances. From this set we selected the 'usable' ones. The criterion was that an utterance

should have a valid phonetic transcription for all words, and that it did not contain hesitations, self-repairs, stutters, heavy background noise or meaningless speech. These assessments are available from the Polyphone database annotations. Based on these criteria we end up with 18,538 phonetically rich sentences for training.

2.1. Training data from 'short utterances'

Short utterances were defined as utterances with no more than three words.

Published Polyphone

The Polyphone corpus contains 44 items (e.g. digits, ZIP codes, times, phonetically rich sentences and application words) spoken by over 5000 speakers. Application words are typically single word utterances or short commands for use in speech driven services (e.g. 'connect', 'open e-mail', etc.). Each caller read four application words. The callers also read two city names (large cities in the Netherlands, and capital cities of large countries). Speakers were asked a questions that elicited extemporaneous expressions of city names[1], viz. "In which cities did you grow up?". Table 1 summarizes the data in terms of utterance length. Most of the multi-word expressions for city names derive from the extemporaneous answers to the question 'Where did you grow up?'.

	application words	city names
# items	20,200	15,150
'clean'	19,688	14,402
# unique items	1375	2311
1 word / utterance	18,692 (94.9%)	10,158 (70.5%)
2 words / utterance	615 (3.1%)	1919 (13.3%)
3 words / utterance	281 (1.4%)	1191 (8.3%)
> 3 words / utterance	100 (0.6%)	1134 (7.9%)
max utterance length	13	17

Table 1: Distribution of utterance lengths of two 'Short Utterance' categories in Polyphone

Polyphone Hidden Items

All speakers in the Polyphone corpus were asked to say their surname, to say their full address. Each item was individually prompted. Thus, the prompting style is very similar to what one would expect in a DA service. The city names are set apart as test set. Thus, we keep 5092 surnames and 5092 street names from which we can select our training material (cf. Table 2). The long utterances are almost exclusively due to spontaneous spelling of the surname and/or street name; letter names are counted as words[2].

	streets	names
# items	5,092	5,092
'clean'	19,688	4,988
# unique items	4,414	3,818
1 word / utterance	3,3,08 (66.9%)	3,263 (65,4%)
2 words / utterance	989 (20.0%)	955 (19.1%)
3 words / utterance	289 (1.4%)	441 (8.8%)
> 3 words / utterance	361 (7.3%)	329 (6.6%)
max utterance length	19	16

Table 2: Distribution of utterance lengths in two Polyphone Hidden items used in this study.

From the total set of potentially useful short utterances from the public and hidden parts of Polyphone, 42,101 utterances were eventually selected for acoustic model training. All utterances are 'clean' (i.e., do not contain hesitations or a high background noise level) and for all words in these utterances valid phonetic transcriptions are available in the Polyphone lexicon and Onomastica for the items from the Hidden part.

2.3 Test corpora

Four test corpora were used, two containing only city name expressions, one comprising phonetically rich sentences and one corpus with utterances from a train time-table application.

City names

The first test corpus comprises the valid expressions of the city of residence of the speakers in Polyphone (i.e., the first item in the Hidden part). Only responses were selected that contained a city name pronounced in isolation (although audible hesitation signals like 'ehr' preceding the city name were allowed). Other disfluencies like ('Amst… Amsterdam') and connected speech responses (e.g. 'I live in Amsterdam') were excluded from the test corpus.

The speakers in Polyphone said the name of the city where they live. In a real-life DA service callers will often ask for a city name other than their own. This could lead to differences in pronunciation. To test investigate this we constructed a second test corpus, recorded in a Wizard of Oz DA service. The selection criteria for including a token in the test set were the same as for the Hidden part of Polyphone. The two test sets for city names are summarized in Table 3. The test corpus comprising the city names from the Polyphone Private Items will be referred to as the PPI corpus; the corpus taken from the DA recordings will be referred to as DDAC2000.

[1] The purpose of asking this question was to gather information on the regional pronunciation variant that would best characterise this speaker.

[2] This part of the recordings cannot be made publicly available because of restrictions imposed by the Dutch data protection Authority., to protect the speakers' privacy.

Phonetically rich sentences

To verify the potential loss of generality of the acoustic models trained on the short utterances and city names, we performed experiments on a corpus of phonetically rich sentences. From the official test part of the Polyphone database 2713 'clean' sentences were selected. These sentences are similar to the sentences in the training corpus in all respects but one: there is no overlap between the speakers in the training and the test corpus.

Corpus	# utterances	# unique city names	max. # of tokens / city-name
Polyphone Private Items (PPI)	4784	963	205
DDAC2000[3]	8973	875	711

Table 3: Details city name test corpora..

Train timetable corpus

This corpus comprises the answers to the first question in the dialog from the operational train timetable information system in the Netherlands. The corpus contains 1,197 utterances. The average utterance length is 4.8 words[4].

3. DESIGN OF THE EXPERIMENTS

3.1 Acoustic model training

With the corpora described in Sections 2.1 - 2.2 two sets of context dependent models were trained for 37 Dutch phones. For transient background noise, like closing doors and passing cars, and speaker sounds, like coughing, we trained two three state HMMs. A third set of context dependent models were trained on application specific data, described in this section.

General purpose models

Despite the fact that the total number of phonetically rich sentences is smaller than the number of short utterances (18.538 versus 42.101) the total number of phoneme symbols in the canonical transcriptions is much smaller in the short utterance corpus (342,521 versus 849,508). Thus, there is much more data to train the models on the phonetically rich sentences.

Moreover, the phonetically rich sentences were designed to ensure that even the least frequent phonemes occurred in each set of five sentences. No such precautions were applied in designing and selecting the short utterances. Yet, it appears that – with the exception of the least frequent phoneme /2:/– the relative frequencies of the phonemes are very similar in the two training corpora (Spearman ρ = .96). In the design of the phonetically rich sentences (and in the short utterances) no explicit attempts have been made to obtain uniform presence of left and right contexts of the phonemes. Still, the ranking of the contexts in both corpora is highly similar. For most phonemes the rank order correlation for the left and right hand phoneme context (Spearman ρ) exceeded 0.7. Thus, the most important difference between the two corpora for training general purpose sub-word models is the overall size of the corpora.

Application specific models

The city name training samples were selected from the database recorded in the Wizard of Oz DA system mentioned in section 2.3. The selection criteria were the same as for the test set, e.g. only city names spoken in isolation were selected, with one potentially important exception: the training corpus contained a substantial number of "I don't know" utterances[5]. The resulting corpus comprises 24,559 utterances, with a total of 194,933 phoneme symbols in the canonical transcriptions. Because the /S/ phoneme occurred only once in the material, context dependent model for only 36 Dutch phone were trained.

3.2 Design of the tests

For the four test three lexicons and three language models were constructed.

City names

The same lexicon of city names and the same unigram language model were used for the two experiments with city name recognition. The lexicon comprised all 2300 city names in the ZIP code directory of the Netherlands. Alternative pronunciations for some cities (like *den haag* for *'s_gravenhage*) were added to the lexicon. If a city name consists of multiple words (like *den haag*). the individual words were concatenated with underscores. The resulting lexicon contains 2373 entries.

The language model for the city name experiments is based on the a-priori probability of the name, estimated from the number of streets in a city as listed in the telephone directory. A language model thus obtained is sub-optimal for both test corpora. In Polyphone there is a bias towards smaller towns in less densely populated parts of the country. In the DA corpus there is a reverse kind of bias: the big cities where most of the economic activity is concentrated are over-represented. Estimating uni grams from the number of streets seems to be a viable compromise.

Phonetically rich sentences

The lexicon for the test with the phonetically rich sentences contained all words which occur in the transcriptions. For the sake of this experiment homophone heterographs were mapped

[3] Acronym for Dutch Directory Assistance Corpus 2000.

[4] The minimum expected utterance length is four words ('from A to B'). Some speakers add 'fillers', like 'I want to travel from A to B'. Yet other speakers also add information about the date and time of the prospective journey.

[5] A substantial proportion of callers to the DA service request premium rate numbers of companies or organisations. Often, the caller does not know where the company or organisation is located.

Model	PPI Cities	DDAC2000 Cities	Phonetically Rich Sentences	Train Timetable service
Short Utterances	7.8 % / 6.8 %	8.4 % / 7.7 %	31.3 % / 77.8 %	13.0 % / 29.8 %
Phonetically rich sentences	13.8 % / 12.06 %	12.5 % / 11.5 %	18.4 % / 60.7 %	13.1 % / 30.9 %
Application Specific	12.9 % / 11.7 %	7.7 % / 7.2 %	45.6 % / 88.0 %	15.5 % / 33.2 %

Table 4: Word Error Rates / Sentence Error Rates for triphone models

onto a single uniform spelling. This was done in order not to obscure a comparison of acoustic models with errors that can only be resolved by means of a the language model. The resulting lexicon contained 5,644 words.

The unigram and bigram language model for the phonetically rich sentences was trained on the prompting texts for the sentences in the training corpus. Therefore, the prompt texts of the sentences that were discarded for acoustic reasons were included in the training of the LM. The test set perplexity was rather high, viz.82.95.

Train timetable utterances

The lexicon for this task contains all valid railway station names in the Netherlands and all plausible 'filler words' that might be used by the customers. The number of entries is 1542. Because the recordings come from operational usage of the system, the test corpus contains some utterances which contain one or more out-of-vocabulary words (2% of the corpus). A language model was trained on an independent corpus with 4667 answers to the same dialog question.

4. RESULTS AND DISCUSSION

The word error rates and sentence error rates of all experiments are summarized in Table 4. For both city name recognition tasks the models trained on short utterances and the application specific models perform better than models trained on phonetically rich sentences. However, there is a clear performance difference between the two city name tasks for the application specific models. This can possibly be explained by the distribution of the cities in the test corpora. The cumulative histograms of both corpora show that the curve is much steeper for the DDAC2000 corpus than for the PPI corpus: 50 city names cover 60% of the DDAC2000 corpus, while 107 names are needed for the same coverage in the PPI Cities. Thus, the application specific models are heavily biased to phonetic contexts in the most frequent city names. The mismatch with cities from the PPI Cities cause a performance decrease compared with the matched case. This suggests that it is difficult to generalize these application specific models to even a closely related task. So far we cannot explain why the models from the short utterances perform better for the PPI Cities than for the DDAC2000 cities. Explanations that invoke a greater similarity between subsets from Polyphone are refuted by the fact that the models from the phonetically rich sentences show the reverse trend.

The models trained on phonetically rich sentences are the clear winners for the matched test corpus. Thus, it appears that these sentences have acoustic properties which are not modeled adequately by the short utterances, nor by the application specific models. In any case, the results show a major impact of speaking (or perhaps better: reading) style on ASR performance.

For the VIOS corpus the differences between the three recognizers are relatively small. The almost competitive performance of the application specific models can at least in part be explained by that fact that the same names (the big cities in the country) are very frequent in both the training and the test corpus. The finding that the models from the short utterances are slightly better than the models from phonetically rich sentences, despite the fact that the latter provide three times as much speech for training, suggests that either the acoustic properties of speech sounds in long read sentences are very different from the sounds in short extemporaneous utterances, or that the training procedure had difficulty in aligning the speech with the phonemes in the transcriptions. Further research is needed to clarify this issue.

In our experiment we have seen little positive effect of training a complete set of sub-word models on speech recorded in the application. In [3] it is shown that inclusion of a small number of application specific sub-word units does improve performance on a city name recognition task. Including an even smaller number of whole word models for the largest cities improves performance even further.

5. CONCLUSIONS

The results in this paper show a dramatic impact of speech style on ASR performance, at least for styles that are very different: reading fairly long sentences versus reading or saying words or short commands. In conclusion we can say that short phonetically rich utterances, even if they are read, seem to be the best possible source of data to train acoustic models. The gain to be had from training acoustic models on application specific data is not very large. It may not justify the additional costs of maintaining multiple sets of models

5. REFERENCES

[1] L Boves, D. Jouvet, J. Sienel, R. de Mori, F. Béchet, L. Fissore, and P. Laface (2000) "ASR for automatic directory assistance: the SMADA project", Proc. ASR2000, Paris, Sept. 2000.

[2] E. den Os, T. Boogaart, L. Boves, and E. Klabbers (1995) "The Dutch Polyphone corpus", Proc. Eurospeech-95, pp. 825-828.

[3] Béchet, F., den Os, E., Boves, L. And Sienel, J. (2000) "Introduction to the ISTHLT project Speech-driven Multimodal Automatic Directory Assistance (SMADA)", Proc. ICSLP2000.