Institute of Phonetic Sciences
University of Amsterdam
Proceedings 19 (1995), 126-134.

PHONEME-BASED AUTOMATIC SPEECH RECOGNITION:
towards a demonstrator for information retrieval, using Dutch hi-fi speech

*Ditta H.Beun, Louis C.W. Pols, and Hans Kloosterman**

Abstract

Phoneme-based speech recognition for the Dutch language, using hi-fi recorded sentences as input, was investigated in this Master Thesis. This study was performed in the framework of a project called Talking Heads, in which KPN Research and TNO-TPD are involved. In that project it will be studied whether integration of automatic speech recognition and information retrieval can be combined in a single system.
A speech database was collected, and context-independent phoneme models were trained and tested. Two topologies, use of a bigram, and several parameter settings were investigated. The best phoneme recognition score achieved, was 52.46%. The models and parameter settings used to obtain this score were applied for exploratory purposes in the Talking Heads project to test a database containing 10,500 words. The phoneme recognition strings thus obtained, served as input for an Information Retrieval System. The models were applied in a demonstrator.

1 Introduction

Growth in the need for information services requires eminent information retrieval systems, capable to retrieve information efficiently and effectively. Such systems using Automatic Speech Recognition (ASR) meet with situations in which speech input is required or desired. Applications for instance in public areas like train stations, airports or tourist information centres might serve the customer with answers to their spoken query. Physically handicapped or elderly people might also be able to access services in a more natural way, since the use of a keyboard is not required. Investigation in services providing speaker-independent speech recognition combined with information retrieval systems deserves much attention.

1.1 Objectives

As part of a project called Talking Heads, a master thesis (Beun, 1995), of which the main results are presented in this paper, was concerned with training and testing Dutch phoneme models, as recognition units of a speaker-independent recognition system, using hi-fi microphone speech as input. In realising this main objective, several activities were performed:

* Hans Kloosterman is employed by KPN Research.

1. Collecting a speech database:
   Collection of a new database required careful preparations. With the Talking Heads project in mind, this concerned defining corpus specifications (such as linguistic content), as well as selection of number and type of speakers and optimal settings for the recordings. How the actual recording should take place was also considered. A final activity in collecting a new database involved editing all the speech files.
2. Training phoneme models:
   This section involved preparing the speech recordings for training and testing. Decisions about what models required training were made. As a final step several iterations were run to model all phonemes.
3. Testing:
   In this section it had to be determined what models perform best. These models then served as a reference in performing off-line tests for the Talking Heads project.

1.2 Automatic Speech recognition

Automatic Speech Recognition (ASR) in this project involves recognising continuous speech by computer. Speech itself is represented as a continuous stream of sound. The listener, the computer in this case, has to segment this stream into discrete units (e.g. phonemes) and to identify these units.

Speech recognition systems nowadays can be based on several methods. Feature extraction, neural nets (artificial intelligence) and pattern recognition are used. In this project Hidden Markov Modelling was used. What Hidden Markov Modelling entails is discussed extensively in Van Alphen's PhD dissertation (1992).

2 The speech database

A new database appropriate for the Talking Heads (TH) project was collected. For us speech corpora are not aims of their own but means to an independently specified purpose (Eagles 1994). Since the purpose of this database is intended to serve similar purposes as the Polyphone corpus, considerations were derived from choices made for the Polyphone corpus (Damhuis et al., 1994).

The speech material recorded, elicited 28 items read by 50 male and 50 female Dutch speakers. Every speaker was prompted with an original set of twenty phonetically rich sentences originating from the Polyphone database (four sets of five sentences). Each set of five sentences was balanced with respect to all possible Dutch phonemes. All speakers were prompted with 3 digit strings and 3 yes/no questions as well. This last material was not used for training and evaluating the phoneme-based recogniser, but was recorded for possible use in other experiments. The yes /no answers, for instance, can be applied to the integrated ASR-IR system, enabling manipulation of an interface with speech.

All recordings were made in an office environment at a 22,050 Hz sample frequency, using a 16 bits linear mono soundblaster card. The microphone (ECM 3003) appropriate for speech was for these recordings attached to a stand, since a microphone stand will also be used in a realistic application. The stand was placed in front of, or next to, the personal computer, minimising the possibility of recording sounds produced by that computer. The speaker was facing the screen.

Recording volume was set at a fixed position. The first item was used by the session leader to determine the speaker's distance to the microphone. All speakers read at

different loudness levels. In order to obtain useful recordings we decided to manipulate the speaker's distance to the microphone. Because, in a real application this can be done more easily than adapting the recording level, unless automatic adjustment is applied.

In accordance with the Polyphone database and the Talking Heads demonstrator, no specific instructions about how a speaker is supposed to read the items was given. We assume that no specific instructions will be given to the user of an integrated Automatic Speech Recognition-Information Retrieval (ASR-IR) system either. Hesitations, mistakes and such things as stutters were corrected.

Besides this database, used for training and testing phoneme models, a database was collected to serve as test database for the TH project. Ten speakers from the above described database (realising a closed test set in terms of using the same speakers, but different recordings for both training and testing) were selected to read 100 words each. The words with various lengths were selected, randomly chosen from a virtual index of 10,500 words from the application domain. Each word was read by five different speakers.

The collected speech corpus, comparable (but not identical) to the Polyphone corpus, consisted of a small amount of training and test material for an automatic speech recogniser. Because the database was collected primarily for an exploratory purpose, a uniform distribution in age, socio-economic status and region was not realised. In order to obtain a balanced database also in this respect, recordings of a larger set of speakers is required. Addressing and recording such a large set was considered too time consuming for this project.

All speech files, were transcribed, representing the spoken utterances in phonetic symbols. A hundred phonetically rich sentences were segmented and labelled for initialising the phoneme models.

The speech corpus, divided in a training and test set, and the TH test set for the IR system are displayed in table 1.

Table 1. Division of the speech corpus in an initialisation, training and test set.

	initialisation set	training set	test set	TH test set
# of male speakers	10	10 + 30	10	5
# of female speakers	10	10 +30	10	5
# of sentences per speaker	5	15, 20	20	100 words
tot. number of sentences	100	1500	400	1000 words

The phoneme was chosen as the unit to be modelled. The database used in this project is not large enough to train word or triphone models. An advantage of the use of phoneme models is that any word can be expressed to and recognised by the system. When only word models are available only the trained words can be recognised.
The speech files were processed to achieve a MFCC (Mel Frequency Cepstrum Coefficients) representation, with a frame length of 23.22 msec. and a frame overlap of 10 msec.

Two topologies were used to train and test phoneme models. The models' topologies are displayed in figure 1.

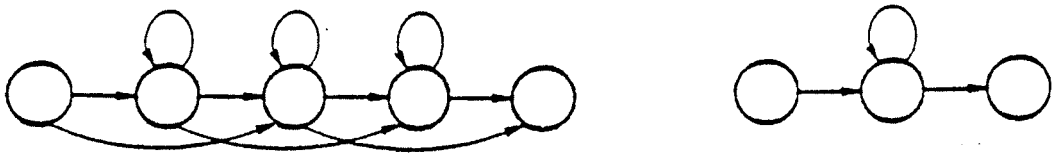

Fig. 1 Topologies of the modelled phonemes.
Topology I (left) consisted of 5 states and allowed 1 skip. The second topology, specified on advice of X. Wang and L. Boves, consisted of 3 states and allowed no skip.

Multiple Gaussian mixtures were applied in the second topology, to account for instance for coarticulation. The covariance matrices of all Gaussian components are diagonal.

3 Experiments and results

Several experiments have been conducted. The first set of experiments concerned choosing the best set of models after several training iterations, using topology I. The experiments were performed with and without the use of a phoneme transition bigram. The second set examined the second topology. Which optional parameter settings, in the used recognition software resulted in highest recognition performance, was examined in the third set of experiments. Experiment 4 investigated what models can be chosen best, using all new parameter settings. Finally a test set of words from the Talking Heads test database was recognised as well. The phoneme output strings of the recognised speech files served as input to the document retrieval system.

Phoneme recognition scores, comparing the phoneme recognition strings as output of the ASR-system to the transcriptions of the speech files, were calculated as follows :

$$\% \; correct = \frac{N-D-S}{N} \times 100\% \qquad (1)$$

$$\% \; accurate = \frac{N-D-S-I}{N} \times 100\% \qquad (2)$$

In these equations N is the total number of phoneme labels, D the number of deletions, S the number of substitutions and I the number of insertions. The accuracy score taking all errors (D,S, and I) into account is chosen as the most representative measure of the recogniser's performance. The best results for all performed experiments are presented in table 2, 3 and 4.

Table 2: Phoneme accuracy scores for experiment 1-4. A phoneme transition bigram was used also in experiments 3 and 4. N= 19250

	analysing:	accuracy score
experiment 1	topology I	41.72 %
	use of a bigram	48.20 %
experiment 2	topology II	41.85 %
	with use of a bigram	48.37 %
experiment 3	parameter settings	51.81 %
experiment 4	best set of models	52.46 %

The additional TH word test could be divided in two sections. In the first section the ASR part of the system is tested. Two symbol sets were used as reference. The HTK-set (table 3) was used for testing the ASR system. The TH set, in which several phonemes were grouped, was developed and applied for matching the ASR phoneme recognition strings to the IR index terms. Both sets are discussed more extensively in Beun's Master thesis (1995).

In the second section the IR system is examined. The words (represented by the TH phoneme strings) were analysed as retrieved successfully either as first element listed as possible correct items, or as one of the first 25 possible correct items. Two retrieval algorithms were applied. TNO's algorithm permitting morpho-syntactic variation, used trigrams as reference. The dynamic algorithm on the other hand uses dynamic mapping, for finding the query term in the TH database index (10,500 words) with monograms (phonemes) as reference. The results are presented in table 3 and 4 respectively.

Table 3: Phoneme accuracy scores for the TH test database.

symbol set	accuracy score
HTK symbols (N=11922)	51.75%.
TH symbols (N=11897)	56.43%.

Table 4: Word scores obtained in testing ASR output (TH phoneme symbol strings) as IR input.

	TNO	dynamic mapping
retrieval in first place	20.7 %	46.2 %
retrieval one of first 25 items	45.2 %	73.9 %

4 Conclusions and discussion

A database, recorded in an office environment, was used successfully for exploratory purposes to train and evaluate an Automatic Speech Recognition system.

The results obtained did not reach the accuracy scores, as for instance achieved by Gauvain et al. (1994). Although our scores can be considered reasonable, several consideration can be made for explaining the differences with the scores obtained by Gauvain et al. These explanations also suggest what can be done to improve recognition performance.

Gauvain et al. (1994) report that they performed speaker-independent large vocabulary speech dictation. They made use of continuous density HMMs with Gaussian mixtures for acoustic modelling and bigram statistics, estimated on newspaper texts for language modelling. Acoustic modelling used cepstrum-based features, context-dependent phone models, phone duration models and sex-dependent models. In table 5 several conditions of Gauvain et al. (1994) are compared with these of our experiments.

Table 5: Conditions of the experiments performed by Gauvain et al. and by Beun (1995).

conditions	Gauvain et al.	Beun (1995)
corpus	WSJ0, WSJ0/WSJ1[1], microphone speech	Microphone, Polyphone -like
test corpus	6 m + 4 f * 10 sentences	10 m + 10 f * 20 sentences
training material	7240 sentences from 42 m + 42 f, 37518 sentences from 284 speakers[1]	40 m + 40 f * 30 sentences = 1600 sentences
modelled unit	context-dependent (triphones)	context-independent
set of phonemes	46	42
number of models trained	493-3306[1]	46[2]
number of mixtures	32	3
gender dependent	yes	no
bigram	yes	yes
phoneme accuracy	66.1%- 83.5%[1]	52.46%

Several considerations can be made about the modelled phonemes.

Incorrect transcription

Discrepancy between normative phonemes and actual pronunciation of phonemes may cause several models to be trained wrongly, because of incorrect transcriptions in this respect. Transcribing all files more carefully, meaning that, for instance, voiceless phonemes sometimes need to be annotated as voiced, or vice versa is time consuming. Since the mentioned variability also remains in real data, grouping or tying phonemes in one model may be considered. Or, as reported by Gauvain et al. use of phonological rules, applied during training may allow for some of the phonological variations in fluent speech.

More than one model for phonemes

Two specific phonemes might be better trained with several models. Phonemes /r/ and /l/ should be modelled as /r/ initial, /r/ final and /l/ initial and /l/ final. The pronunciation of these sounds, especially /r/, in initial or final position is so different that the models currently trained, are not discriminating enough. A phoneme confusion matrix demonstrated that /r/ was frequently confused with various other phonemes. This may have been caused by training this model with both initial and final segments. Confusions may have been made because of the small segments of /r/ in the labelled data. Segmenting /r/ consistently is hard. Segments also contain information of neighbouring phonemes, due to coarticulatory effects.

Context- and gender-dependent phoneme models

The coarticulatory effects bring us to another issue: generalised triphone models or context-dependent models. Context-dependent models are one way to represent coarticulatory effects. Applying models like these will result in better recognition scores

[1] Gauvain et al. trained the 3306 models with WSJ0/WSJ1 si-24 training data.
[2] Including a model for mouth_noise and begin and end silence.

at the cost of greater complexity and the need for more training data. Gauvain et al. used context-dependent models in their recognition experiments. The use of these models may be one of the explanations for better recognition results achieved in their experiments.

Training gender-dependent phoneme models may also be considered. Male and female speakers have sex-related speech characteristics, partly due to physiological and anatomical differences. Cultural factors and sex role stereotypes also play an important role (Eagles 1994). Since differences involve variation in pitch, intensity, overall spectral slope and accuracy of pronunciation, training distinct models for both sexes separately, as was done by Gauvain et al. (1994), may result in higher recognition scores.

Language models at word and phoneme level

Besides improving phoneme models, recognition may also be improved by applying a word network. The use of a word network has several disadvantages though:
- computation time is increased;
- not all words can be expressed to the system;
- all new words imply changing the network.

When an application with a small vocabulary is indicated, the use of a finite state network must be considered.

Several features were used during recognition. The use of a phoneme transition bigram improved recognition performance. Using a bigram with models trained according to topology II, resulted in an average improvement in recognition performance of 15.42%. This was relatively computed, compared to the recognition performance of the models without use of a bigram.

Parameter settings

Experiments with different values for two software parameters manipulating the numbers of deletions and insertions during recognition, indicated that similar accuracy scores can be obtained using different combinations of parameter values. The number of deletions and insertions change with altering values. A set of values was chosen to conduct experiments with. This set may not have been the best set of values. Better accuracy results might have been obtained with another combination of values. Unfortunately it was not possible to test more combinations of values to find the best set of values.

Size of the database

A relatively small database was used to train and test the Dutch phoneme models. Collecting a larger database was not possible in this project. Collection of a larger database is necessary when a robust, speaker-independent phoneme-based, application will be built. For exploratory purpose a small database may suffice. A robust speaker-independent and perhaps sex- and task-independent recogniser requires the use of a large database.

Recognition scores obtained by Gauvain et al. for instance were based on models trained with a large database.

When collecting a large database, the following two points are important :
- enough occurrences of each token (context-independent or context dependent phoneme) to be modelled must be available;
- a uniform division of speakers over characteristics like gender, region of accent and for instance level of education is necessary to train robust models.

Duration modelling

Two different topologies were tested. Use of other topologies may be investigated. Duration, for instance, was not specifically modelled. This may explain why accuracy scores for the TH set were slightly less than obtained with the test set consisting of phonetically rich sentences. The words (of the TH set) expressed in isolation were uttered more carefully. Final /n/'s, for example, are often omitted, or uttered very shortly in sentence context. A durational difference between final /n/'s pronounced in isolated words compared to pronunciation in sentence context, may therefore occur. Besides durational aspects, the use of more mixtures should be investigated. Experiments conducted by Gauvain et al. made use of topologies consisting of 32 mixtures (see also table 5).

Robustness of speech recordings

Several considerations can be made regarding the speech recordings and processing of these recordings, before we applied them for training and testing. Background noises, for instance, were recorded, which fortunately had a low level compared to the spoken sentences. Although speakers read all prompted items at different loudness levels and at different distances to the microphone no specific instructions should be given in this respect. The various loudness levels and microphone distances may be encountered in realistic settings as well. Choice of the right microphone, may minimise recording from specific surrounding noises.

Unfortunately, a few of the prompted items were not recorded completely. The first phonemes were not always recorded, because in a few cases, the speakers started reading too soon. A 100 % correct score could therefore never be obtained for these files. In a real application this problem may be encountered as well. Uttering the query to the system for a second time may result in a better score.

Band pass filtering is another issue that may be indicated. The speech files in this study were not filtered before they were used for training and testing. Valuable speech information is mainly found in frequencies up to 5 kHz (Holmes 1988). Low-pass filtering at 5 kHz will discard all frequencies above 5 kHz as well as unnecessary feature vectors.

When an ASR-IR system is applied in different environments, for example in a train station, different background noises may be expected and are therefore not accounted for in the current phoneme models. Models used in these kind of settings need to be more robust. It needs investigation though whether new models ought to be trained for each new setting or whether one set of models recorded in different settings can be applied perhaps after some adaptation. Training new models for each new purpose is an enormous task. Databases then need to be collected for each new application.

When use will be made in other projects of the database recorded for this study, one must keep in mind that the database is not large enough to train robust ASR systems. A uniform balance over characteristics, except for gender, was not achieved. Besides, the database is not large enough to train generalised triphones.

5 General conclusion

Comparing the experiments conducted in this project to experiments performed by Gauvain et al. suggests that several improvements to speaker-independent recognition can be made. In our opinion integrating an Automatic Speech Recognition system and an Information Retrieval system in one application in an office environment appears to

be possible. The mentioned improvements involve a lot of work and effort though, but will be beneficial in the end.

Acknowledgements

The first author would like to express her gratitude to the speech and language group of KPN Research' department for Systems integrations and Application for Multi-media (SAM), for giving her the opportunity to work on this project at KPN Research.

References

Alphen, P. van (1992): *HMM-based continuous-speech recognition. Systematic evaluation of various system components*, Ph.D. thesis, University of Amsterdam.

Beun, D.H. (1995): *Phoneme-based automatic speech recognition. Development of an ASR system for integration in a demonstrator for information retrieval, using Dutch hi-fi speech.* Master thesis, r&d sv 95-1086, KPN Research, Leidshendam.

Damhuis, M., Boogaart, T. , in 'T Veld, C., Versteijlen, M., Schelvis, W., Bos, L. & Boves, L. (1994): "Creation and analysis of the Dutch Polyphone Corpus", *Proceedings of the International Conference on Spoken Language Processing (ICSLP' 94),Banf.* vol 4: 1803-1806.

Eagles, (work in progress version October 1994): *Spoken language systems, Chapter 2 Spoken Language Corpora.*

Gauvain, J.L., Lamel, L.F., Adda, G. & Adda-Decker, M. (1994): "Speaker-independent continuous speech dictation" *Speech Communication,* **15**: 21-37.

Holmes, J.N. (1988): *Speech synthesis and recognition.* Reprinted in 1993, Chapman & Hall, London, United Kingdom.

HTK. (1992): *Hidden Markov Toolkit Version 1.4 Manual.* Cambridge University Engineering Department, Speech Group, Cambridge.

Wang, X. (in progress). Ph.D. Thesis, University of Amsterdam.