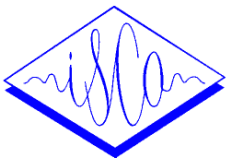

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

THE PHILIPS LARGE-VOCABULARY RECOGNITION SYSTEM FOR AMERICAN ENGLISH, FRENCH AND GERMAN

Christian Dugast, Xavier Aubert, Reinhard Kneser

Philips GmbH Forschungslaboratorien, P.O. Box 1980, D-52021 Aachen, Germany
E-mail: {dugast,aubert,kneser}@pfa.philips.de

ABSTRACT

This paper presents the work done at Philips Research to extend our American-English large-vocabulary continuous speech recognition system to two new languages, French and German. The tasks on which the system is tested are very similar for each language: a speaker-independent, continuous-speech, national newspaper reading task with a high-quality microphone. However, an important factor for German is the extension of the recognition vocabulary from 20k to 64k words to achieve comparable out-of-vocabulary rates. A comparison between the different databases used to train the system for the three languages is made, results are accordingly interpreted and the characteristics of each language are pointed out.

10.21437/Eurospeech.1995-49

1. INTRODUCTION

Aim of the SQALE project (within which this work has been carried out) is to define a framework to make comparisons across different continuous-speech recognition (CSR) systems in the European multilingual environment. The comparison study should try to answer the two questions: how do different systems compare on a given language (i.e. what are their technological means to tackle language-specific problems?), and how do languages compare with each other? The scope of this paper will be restricted to answering the second question: how the Philips system compares on a newspaper reading task between the three languages, American English, French and German.

First, we will present our base-line system. A new feature in our system is the capability of working with pronunciation variants. In the next section, we describe for each language their respective training corpus. An analysis in terms of lexicon coverage, homophone rates, perplexity and average number of phonemes per spoken word for the three languages is provided.

Recognition results are presented with the same baseline system for all three languages. Language specific optimizations are presented and discussed.

2. BASELINE SYSTEM

Architecture of the Recognition System: Recognition is carried out in two passes. The first pass, an integrated search, is based on a left-to-right time-synchronous beam search. Limitations of this first pass are the use of within-word context-dependent (CD) phone models and of a bigram language model. The output of the beam search is either the best bigram-scored sentence or a word-lattice that can be processed in a second pass. During the second pass, the word-lattice is generalized into a graph, on which longer spanned language-models (i.e. 3-grams) may be applied.

Acoustic Analysis: A 63-component feature vector is computed every 10 ms. It consists of 30 log-spectral coefficients. For every second spectral intensity, as well as for the energy, first and second differences are also computed making altogether 63 coefficients. Linear Discriminant Analysis is then computed at the state level, so as to improve state separability. The dimension of the feature vector is then reduced to the 35 most significant components [2].

Acoustic-Phonetic Modelling: The basic units of our acoustic-phonetic models are within-word CD-phones. Each CD-phone is modelled by a Hidden Markov Model. A phoneme consists of three segments. An emission probability density function (mixture of Laplacians) is attached to each segment. In contrast to other systems, the Viterbi criterion is used both in training and recognition. This applies even to the level of mixture components, such that the sum over the component densities is replaced by their maximum [3]. Moreover, only one variance vector is estimated, pooled over all mixtures. State-tying of mixture means [4] has also been implemented [5].

Tree Lexicon: The pronunciation lexicon is arranged as a tree of CD-phones (tree lexicon). The handling of pronunciation variants has been recently introduced in our system [6]. It allows to create automatically a pronunciation-dependent script of all training sentences.

Language Modelling: Word trigram language models are estimated on the same kind of newspaper corpus on which evaluation occurs.

3. DATABASES

The same training and testing conditions have been set up for all languages with the exception of German, where the spoken training corpus was not newspaper reading but the so-called PHONDAT corpus, which

This work has been partly funded by the European Union within the SQALE project, LRE-62-058

consists of a "diphone corpus" and "train time-table inquiries". The heterogeneity between training and testing material for German is increased by the use of four different microphones to record PHONDAT and a fifth one to record the German newspaper sentences on which evaluation has been made (Sennheiser HMD 224 X). Four sites recorded Phondat, each site having used its own microphone: 2 Sennheiser (MKH 20 P48 and MD 412) and 2 condensator microphones, one from AKG, the other from Neumann (U67). This mismatch that makes the recognition task more difficult for the German corpus than for the other languages, is to be taken into account while comparing results across the three languages.

The text corpora used to train the trigram language models are of about the same size for all languages, ranging between 32 and 38 million words.

3.1. American English

The SQALE project decided to use the WSJ0 (Wallstreet Journal) [7] training data (84 speakers) to train the acoustic-phonetic models for American English. Speech is roughly equally balanced in gender (see table 1).

Starting from the Dragon lexicon, the phoneme symbol set has been reduced to 44 units plus silence, compacting all stressed variants of each phoneme to one unique symbol. The lexicon has been corrected to make it more consistent and, in addition, about 11% of the words received one or more pronunciation variants (see table 1).

The non-verbalized 20k WSJ trigram language model from Doug Paul (ARPA evaluation fall 1993) has been used for recognition. It has been trained on 38 million words from the "Wallstreet Journal", a financial newspaper.

3.2. French

The BREF-80 French corpus has been made available by LIMSI to all SQALE partners [8][9]. From the 80 speakers, actually only 76 have been used to train the different systems (BREF-76). The gender distribution is not even, favorizing females (table 1).

The original BREF lexicon from LIMSI has been used for training and testing. The number of phonetic symbols used to transcribe the lexicon is 33, a very small figure compared to both other lexica. The proportion of pronunciation variants for French is smaller than for American English, being 5% (resp. 3.5%) of the overall training vocabulary (resp. recognition vocabulary).

The 20k trigram language model from LIMSI has been used for recognition. It has been trained on a 38 million word corpus extracted from "Le Monde", a French national newspaper.

3.3. German

The acoustic models for German were trained on a subset of the "PHONDAT" database (CD-ROMs 1-4 from PhondatI, CD-ROM 1 from PhondatII). Compared to the other languages, the number of spoken utterances is much higher (factor 2 to 3) and the variety in speakers is also higher (factor 2). But the amount of speech is very similar to what has been used for WSJ0. We may note the huge proportion of silences within PHONDAT (40%,see table 1).

Table 1: General characteristics of the different databases

		A-Engl.	French	German
# Speakers		84	76	157
	males	42	33	75
	females	42	43	82
Speech		12:20	7:55	11:50
(h:m)	males	6:00	3:15	6:20
	females	6:20	4:40	5:30
Silence (h:m)		3:00	2:00	7:30
# Utterances		7240	5063	19125
	males	3586	2226	10031
	females	3654	2837	9094
Lexicon	# units	44	33	49
Training	# words	9083	13873	1717
	# variants	1077	663	-
Test	# words	19979	20000	64142
	# variants	2219	717	-
Language model		WSJ	Le Monde	FR
	# words	38 M	38 M	31.5 M

The Philips lexicon transcribed with 49 phoneme symbols plus the silence did not include any pronunciation variants. It totalizes 1717 (resp. 64142) different words for training (resp. recognition). The phonetic transcriptions have been automatically generated [10].

3.3.1. Description of Phondat I and II

Phondat consists of two parts, a so-called diphone-corpus and a train inquiry corpus. Phondat is the result of a project funded by the German government.

PhondatI, the diphone-corpus, consists of 397 phonetically balanced sentences, the alphabet, the numbers 0–12 and two standard stories, read in one piece. The stories are entitled "Nordwind und Sonne" and "Buttergeschichte". This material was designed in such a way that all possible German phoneme-transitions are covered, including diphones containing word boundaries. This makes 1308 distinct phonemic dyads, each appearing at least once in the material.

Speech was recorded under studio-quality conditions. Two hundred speakers are recorded, 100 of each gender. The median in the age distribution is 30 years. Moderate regional accents were covered by appropriate selection of the speakers. The total of 441 (sentence-equivalent) utterances of the corpus was distributed over the speakers.

The sentences were read from a prompting screen. The speech was recorded at 16 kHz sample rate with a dynamic resolution of 16 bits.

PhondatII consists of 200 utterances from the domain of train inquiries. The texts for the recording material are based upon real spontaneous dialogues

(University of Regensburg). From these dialogues, 200 sentences were selected.

Sixteen speakers have been recorded (10 males/6 females). The recording conditions were the same as for the PhondatI corpus.

For all speech files, orthographic transcriptions are available.

3.3.2. The Frankfurter-Rundschau Corpus

The corpus from which the trigram language model has been evaluated for the German language has been in the meantime made available through ECI (European Corpus Initiative) and LDC. It stems from a German newspaper, the Frankfurter Rundschau (FR). Even though it is a nationwide newspaper, 40% of its content is local news (around the city of Frankfurt). The corpus is made of all articles published during the period July 1992 through March 1993 and has a total of 31.5 million words (1.5 million sentences). The corpus has been processed at Philips so as to treat as homogeneously as possible abbreviations, punctuation, numbers and word capitalization.

The publication period corresponding to the last month (March 1993) has been separated from the corpus to serve as test set. The 800k words it covers were used to evaluate perplexities and Out-Of-Vocabulary (OOV) rates. Hence, 30.8 million words were left for language model training.

German is a highly inflected language. For example, the number of different words occurring in the 31.5 million word corpus is greater than 600k, which is to be compared to the 150k (resp. 280k) different words extracted from the American "Wall Street Journal" (resp. French "Le Monde") 38 million word corpus. A 20k vocabulary would then have for German an OOV rate of nearly 12%, to be compared with the less than 2% OOV rates observed for the other languages. It was therefore decided to augment the base recognition vocabulary from 20k for American English and French to 64k for German, for which (on the test set "March 1993") an OOV rate of 6% could still be observed.

Perplexities on the above defined test set are a factor of 2 higher compared to what could be observed for both other languages: 430 for bigram, 336 for trigram language models.

4. LANGUAGE-SPECIFIC FEATURES

When pronunciation variants are available for a language, the training corpus is newly transcribed according to the optional phonetic transcriptions. It generates what we call a real script. A gender-independent LDA is then evaluated on this real training script. Within-word triphones and diphones are also selected for each language from their respective real training script. This selection is based on a minimum number of occurrences. For gender-dependent modelling, the selection has been made along the respective gender-dependent scripts.

4.1. American English

For gender-dependent modelling, the triphone (resp. diphone) occurrence threshold (above which a new model is trained) has been fixed at 35 (resp. 50) which gave a total of about 2000 context-dependent models (or 6000 states) for each gender. For gender-independent modelling, the thresholds have been set higher, 40 for triphones and 60 for diphones giving a total of about 2400 context-dependent models (7200 states). After state-tying, the total number of states to model has been reduced down to 2179 for female, 2869 for male and 3157 for gender-independent modelling. For each tied state, a mixture of Laplacian components has been trained with an average of 30 components per mixture for gender-dependent modelling and 36 for gender-independent modelling.

4.2. French

A characteristic of the French language is the use of liaisons. A liaison is the insertion of a consonant at the end of a word (corresponding to the last letter of the word) if the word coming next to it begins with a vowel. The liaison-rules suffer from numerous exceptions and are optionally used in real life. The words that can be followed by a liaison phoneme have been marked in the lexicon. The real training script, for French, has then been produced with respect to pronunciation variants and liaisons [6]. Within-word CD phones have been selected accordingly. For gender-dependent models the thresholds have been fixed at 35 triphone and 50 diphone occurrences giving a total of about 1400 CD models (4200 states) for each gender. For gender-independent modelling, the thresholds have been set higher, 65 for triphones and 85 for diphones giving a total of about 1400 context-dependent models. After state-tying, the total number of states to model has been reduced down to 2116 for female, 2186 for male and 2456 for gender-independent modelling. For each tied state, a mixture of Laplacian components has been trained with an average of 30 components per mixture for gender-dependent modelling and 39 for gender-independent modelling.

Table 2: Comparative results on three languages, evaluation set May 1995

	A-Engl.	French	German
Word-error rate	**14.7%**	**16.1%**	**19.7%**
Vocabulary size	20k	20k	64k
bigram perplexity	191	172	359
trigram perplexity	131	116	271
Av. # phones/word	4.1	3.5	4.9
OOV rates	1.5%	1.7%	1.9%
Av. # Errors/OOV	1.6	1.7	1.5
Tri. + Diph. coverage	97.4%	96.4%	83.8%
Err. due to homoph.	3%	25%	3%
# Speakers	20	20	20
# Spoken words	3415	3357	3303

4.3. German

Gender-independent models have been set-up for 1244 triphones and diphones, the total number of states after tying being reduced to 2858. The reduction factor in the number of states is 1.3 compared to a factor 3 for American English. On average, there are about 40 densities per acoustic state. The gender-dependent models have been estimated for about 1300 contexts that correspond approximately to 3000 tied states, each with an average of 30 acoustic densities. State-tying happend to be, for German, not as rewarding as for the other languages.

5. RESULTS

Results are discussed based on the figures observed during the SQALE evaluation fall in May 1995. The discussion refers to table 2. American English will be taken as the standard language to compare with.

5.1. French

For French, the handling of liaisons during training and recognition played a significant role as it brought about 10% relative improvement in the word error rate [6].

Looking at table 2 for French, there is an impressingly high percentage of errors (25%) due to homophones. A trigram language model does not have enough information to decide for the correct written form. Example: "Il s' était tu" pronounced the same way as the plural sentence "ils s'étaient tus", where the "ils" refers to something elsewhere not caught by the trigram.

Another characteristic of the French language, is the high percentage of very short words (table 3), which might give bad surprises when an OOV occurs like "s'épanouissait" recognized as "c'est pas nous oui c'est" which is grammatically correct (and makes sense). However, the average number of errors per OOV is not much higher than that on the WSJ task.

5.2. German

There are two points of importance to notice before interpreting the results on the German corpus. First, the OOV rate as well as the perplexities on the evaluation set are low bounded values that are not representative of the difficulty of the task (see section 3.3.2). Second, as already mentioned, the microphone used to record the evaluation set was different to any of the 4 microphones used to record Phondat, the training corpus. And no attempt, besides long term normalization, has been made to alleviate this mismatch with microphone-adaptation procedures. But other factors show the difficulties of the German task: First, the extraordinarily high perplexity compared to the 2 other languages, is probably not sufficiently compensated for by the average number of phones per word. Actually, German has here, on average, 1 phoneme more per word than American English. But this additional phoneme is often a flection like 'em', 'en', 'er' or 'es', which are highly confusable.

Second, the low tri- and diphone coverage on the test set recalls that the corpus used to train the acoustic models (Phondat) is very different to the testing task and is probably not complete enough.

Very surprising for German is the rather low average number of errors per OOV, although about 20% of the OOV have been recognized in a split form, producing 2 errors each (i.e. "Künstler Name" recognized in place of "Künstlername"). One reason might be the low percentage of words having less than 3 phonemes (see table 3).

Table 3: Percentage of words transcribed with 1 or 2 phonemes

	A-Engl.	French	German
1 phone words	2%	16%	0.4%
2 phone words	24%	27%	16%
1 + 2 phone words	26%	43%	16%

6. REFERENCES

[1] X. Aubert, H. Ney: "Large Vocabulary Continuous Speech Recognition using Word Graphs", Proc. ICASSP-95, Detroit, pp. 49-52, MI, 1995.

[2] R. Haeb-Umbach, H. Ney: "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition", Proc. ICASSP-92, San Francisco, CA, pp I13-I16, 1992.

[3] H. Ney: "Modeling and Search in Continuous Speech Recognition", Proc. EUROSPEECH-93, Berlin, pp. 491-500, Sept. 1993.

[4] S.J. Young, P. C. Woodland: "The Use of State-Tying in Continuous Speech Recognition" Proc. EUROSPEECH-93, Vol. 3, pp. 2203-2206, Berlin, Germany, Sept. 1993.

[5] C. Dugast, P. Beyerlein, R. Haeb–Umbach: "Application of Clustering Techniques to Mixture Density Modelling for Continuous-Speech Recognition", Proc. ICASSP-95, Detroit, MI, 1995.

[6] X. Aubert, C. Dugast: "Improved Acoustic-phonetic Modeling in Philips' Dictation System by Handling Liaisons and Multiple Pronunciations", somewhere else in these Proceedings, EUROSPEECH-95, Madrid, Sept. 1995.

[7] D.B. Paul & J.M. Baker: "The design for the Wall Street Journal-based CSR corpus", In PROC. FIFTH DARPA SPEECH AND NATURAL LANGUAGE WORKSHOP, pages 357–362. DARPA, Morgan Kaufmann Publishers, Inc., 1992.

[8] J.L. Gauvain, L.F. Lamel, M. Eskenazi: "Design Considerations and Text Selection for BREF, a large French read-speech corpus," Proc. ICSLP-90, Kobe, Japan (1990).

[9] L.F. Lamel, J.L. Gauvain, M. Eskenazi: "BREF, a Large Vocabulary Spoken Corpus for French," Proc. EUROSPEECH-91, Genova, pp. 505-508, Italy (1991).

[10] S. Besling, "A Statistical System for Grapheme-to-Phoneme Conversion", Proceedings of the Tenth Annual Conference of the UW Centre for the New OED and Text Research, pp. 5-13, October 20-21, Waterloo, Ontario, Canada, 1994.