HMM/MLP Hybrid Speech Recognizer for the Portuguese Telephone SpeechDat Corpus

Astrid Hagen[1] and João P. Neto[1,2]

[1] L^{2}F Spoken Language Systems Lab
INESC-ID, Rua Alves Redol 9, Lisbon, Portugal
{Astrid.Hagen,Joao.Neto}@l2f.inesc-id.pt
[2] Instituto Superior Técnico, Portugal

Abstract. In this article, we describe an automatic speech recognizer developed for Portuguese telephone speech. For this, we employed the Portuguese SpeechDat database which will be described in detail, giving its recording conditions, speaker characteristics and contents categories. The automatic recognizer is a state-of-the-art HMM/MLP hybrid system employing different kinds of robust acoustic features. Training and testing was carried out on the clean digits and numbers part of the database. The recognition results show competitive performance to similar systems developed for other languages.

1 Introduction

SpeechDat is a series of projects to collect speech data, which are funded by the European Union[1]. The aim of the SpeechDat data collections is to establish spoken language resources for the development of voice operated teleservices and speech interfaces. Spoken language resources are speech databases including annotations, pronunciation lexica, and material for the creation of language models, which are needed for the development and use of speech recognition (and synthesis) technology.

During the recording of speech data, the type of microphone (and its position) can already drastically influence the speech signal. Even more importantly, recordings over the telephone line introduce severe distortions due to the telephony transmission channel. With the large variety of telephone gadgets and transmission line characteristics which exist today, such attenuation distortions are hard to predict. The limited bandwidth of the transmission channel of 200/300-3200/3400 Hz additionally restricts the quality of the speech presentation. For these reasons, utilizing a speech recognizer over the telephone line which had been trained on data not recorded over a telephone line or sometimes even only on a different transmission channel can lead to a severe degradation in recognition accuracy. Thus, the availability of large, telephone-recorded databases is important to the research and development of competitive, state-of-the-art speech recognizers for teleservice applications. Such a database is now

[1] http://www.speechdat.org

N.J. Mamede et al. (Eds.): PROPOR 2003, LNAI 2721, pp. 126–134, 2003.

also available for (European) Portuguese and we present its main characteristics in the following section. In Sect. 3, we describe our HMM/MLP hybrid recognizer developed on this database and give first test results in Sect. 4.

2 Database Description

The Portuguese SpeechDat database[2] has been developed within the SpeechDat project to address current and future requirements in the field of telecommunication, spoken language technology and research [8]. It has been recorded in two phases over the public telephone network involving a large set of speakers, recording conditions and tasks. In the first phase (SpeechDat 1), there were 1,000 speakers involved, in the second phase (SpeechDat 2) 4,000 speakers.

The Portuguese SpeechDat database was collected by Portugal Telecom via digital line (ISDN). The design and post-processing of the database, including linguistic annotation, was carried out by INESC[3]. The design of the collection platform and the recording of the speech data itself were effectuated by INESC-TEL. Speech signals are recorded at 8kHz, 8-bit A-law format. The database comprises 14 CDs (3 CDs for SpeechDat 1 and 11 CDs for SpeechDat 2).

2.1 Call Description

Each telephone call included in the database comprises two parts: a first part in which the caller was asked to provide spontaneous answers to nine questions (cf. Table 1), and a second part in which he/she should read a list of 33 items.

The answers about "name" and "telephone number" were not included in the CDs due to privacy restrictions.

Table 1. The nine SpeechDat categories used in the first part of each call to produce spontaneous speech

Está pronto a comea̧r?	Are you ready to start?
Por favor, diga o seu nome.	Please say your name.
Diga o seu número de telefone.	Say your telephone number.
Qual a data do seu nascimento?	What is your birthday?
Qual a cidade (ou distrito) em que passou a maior parte da sua infância?	In which city (or district) have you spent the largest part of your childhood?
É do sexo masculino?	Are you male?
Está a usar um telemóvel?	Are you using a mobile phone?
Está a usar um telefone sem fios	Are you using a cordless phone?
Que horas são?	What time is it?

[2] http://www.l2f.inesc.pt/resources/spdat/speechdat.html

[3] http://www.l2f.inesc-id.pt

After responding to the spontaneous part, the caller is asked to read the sheet number and is then prompted with 32 items to read, which correspond to this sheet. The sheet number consists of a 4-digit number which the caller is asked to read as a digit sequence. Some callers, however, did not stick to this guideline but preferred to read it as a natural number.

The prompted items contain: an isolated digit, three natural numbers, a credit card, a telephone and a PIN number, two money amounts, two dates, one time indication, six application words, three spelled words, three word spotting phrases and nine phonetically rich sentences.

1. The isolated digits are the 10 digits *zero* (zero) to *nove* (nine), and the female forms of "one" and "two": *uma* and *duas*.
2. The natural numbers include the digits and all multiples of 10 and 100, *mil* and the word *e* (and).
3. The credit card numbers consist of 4 times 4 digits, e.g. 8654 3374 1250 6017, whereas the telephone numbers comprise 6 to 7 digits, approximately corresponding to the distribution of the telephone numbers in Portugal at that time (40% with 6 digits, 60% with 7 digits).
4. The money amounts contain small (< 10,000$00) and large (> 10,000$00) amounts, as well as the Portuguese words for the former Portuguese currency *escudos*, *centavos* (cents) and *contos* (1,000 escudos).
5. The time phrases include five different types:
   - **meio-dia** (midday),
   - (*meia-noite*) **e um quarto**, (a quarter past midnight),
   - (*uma*) **e meia** (half past one),
   - **um quarto para** (*meia-noite*) (a quarter to midnight), and
   - (*duas*) **e um|dois|três...**,

   as well as the following days: *ontem* (yesterday), *hoje* (today), *amanhã* (tomorrow).
6. The dates have the following form: <day-of-the-week>, <day-of-the-month> of <month> of <year>, such as *Monday, 1st of January, 1996*. As the Portuguese days (*segunda-feira* (Monday), *terça-feira* (Tuesday), ...) are often pronounced without the indication *feira* (day), the read dates in analogue form have been generated with and without this word.
7. 60 different application words were used, such as *telefonar* (to call), *ajuda* (help) or *agenda* (agenda), one of which was included in each sheet.
8. The words to be spelled were the most frequent proper names in Portuguese, and were presented in the following manner: C,O,N,C,E,I,Ç,Ã,O. Most often diacritics were ignored while spelling. When not ignored the vocabulary typically used for specifying diacritics was: *til* (tilde), *cedilha* (cedilla), *cedilhado* (with tilde), *acento* (accent), *agudo* (as in Á), *grave* (as in À), *circunflexo* (as in Â), *com* (with). Other expressions though can also be found such as "*C de cão*", "*Q nove*" or "*Q de aste*", or reading of the word by pronouncing each syllable separately instead of actually spelling.
9. For the word spotting phrases, 300 different sentences containing the application words were designed. Three of these sentences where included in each prompt sheet.

10. The phonetically rich sentences were created in such a manner as to include in each sentence at least two examples of each phone and as many different triphones as possible.

These items were presented in an alternating fashion in order to avoid fatigue of the speaker.

2.2 Speaker Recruitment and Characteristics

The speakers were recruited among the employees of Portugal Telecom. As the company has a wide geographical coverage, a good representation of many regional accents was guaranteed. The distribution of male and female speakers amounts to 45% male and 55% female speakers, in the age of fourteen years old to older than sixty. Most speakers are from continental Portugal, but some speakers from other areas, such as the Açores (28 speakers), Madeira (8), Africa (32), Macau (1) and others (9) were also included. Most of the speakers born in Africa (Angola, Moçambique, Guiné, Cabo Verde, São Tomé and Principe) have been living in continental Portugal for many years so that their original accents have been reduced.

2.3 Database Annotation

For each available speech signal file exist a corresponding description file and a comments file, in which e.g. the gender and origin of the speaker are stored as well as the transcription of what was uttered. The utterances were annotated on the word level by three experienced annotators.

The speech data of the first phase (SpeechDat 1) was labeled for start and end point of 13 different noise cases. In the second phase (SpeechDat 2) the noise cases were merged into four remaining classes and roughly marked in every utterance. These four noise classes are:

- *filled pauses*: [ah], [eh], [hum], . . . ,
- *speaker noise*: loud breath intake, throat clearing, coughing, . . . ,
- *non-speaker noise*: line noise, radio playing, background voices, . . . ,
- *other evens*: truncated or mispronounced words, background noise, ...

The lexicon of the entire database consists of approximately 19,744 words. The broad phonetic transcription was carried out using the SAMPA symbols. Only one pronunciation is indicated per word and corresponds to the pronunciation used in the region of Lisbon and usually in the media. The transcription was automatically generated and then manually corrected by a phonetician.

3 ASR System Setup

In this section, we describe the training and test sets, the acoustic modeling used in this work, as well as the vocabulary and language models employed.

3.1 Training and Test Set Definition

Given the size of the corpus with its many different contents categories, we decided to concentrate on the the digits and numbers part of the database, more precisely the categories B1, C1–4 and I1 described in Table 2. These categories are especially important to such application domains as credit card and account number validation, automated dialing, user identification via PIN codes, and others.

In this work, sentences of which the transcription contained markers for truncated speech, mispronunciations, or unintelligible speech, or noise markers for speaker noise or line/background noise were disregarded and only the clean utterances were used. The training and cross-validation set for the digits and numbers part of the SpeechDat 1 and 2 database comprises 9981 clean[4] utterances (13h 24min of speech), roughly equally distributed in terms of utterances over the six numbers categories as shown in Table 3. The test set consists of 929 clean utterances (1h 14min of speech), distributed as shown in Table 3. The sets correspond to the defined partitioning of the speakers into training and test set as given on the SpeechDat CDs, so that each speaker was only used in either of the sets.

Table 2. Illustration of the digits and numbers classes of the SpeechDat database

Class ID	Class contents	Example To read	As has been read
B1	10 isolated digits	0965423871	”zero nove seis cinco quatro dois três oito . . . ”
C1	Sheet number	33546	”três três cinco quatro seis”
C2	Telephone number	090981696	”zero noventa nove oito um seis nove seis”
C3	Credit card number	4585 4567 . . .	”quatro mil quinhentos e oitenta e ...”
C4	PIN code	159.160	”cento e cinquenta e nove mil cento e . . . ”
I1	1 isolated digit	6	”seis”

Table 3. Distribution of the utterances in the training and cross-validation set (left) and in the test set (right) over the six classes of the SpeechDat database used here

	Training	Test
B1:	1461	110
C1:	1606	144
C2:	1770	179
C3:	1621	180
C4:	1566	117
I1:	1957	199
SUM	9981	929

[4] “Clean” here signifies no speaker or background noise though moderate noise introduced by the telephone network is a natural consequence of the recording conditions.

3.2 Acoustic Modeling

An alignment was created with Gaussian models, using flat start, and then refined with the MLPs, using the clean utterances of the better labeled first part of the corpus (SpeechDat 1). These MLPs were then used to align the clean utterances of the second part (SpeechDat 2). The whole set of training utterances was then re-aligned several times.

The use of reliable features is a key issue in the design of an automatic speech recognition system. In this work we investigate 3 feature streams (i) 12 PLP cepstra and the log energy, (ii) 12 RASTA-PLP cepstra and the log energy, and (iii) 28 Modulation Spectrogram (MSG) features, extracted on windows of 20ms with a frame shift of 10ms. The PLP and RASTA-PLP features are extracted from the auditory spectrum after filtering the power spectrum with trapezoidally shaped filters applied at roughly Bark intervals, equal loudness pre-emphasis and cube root compression. The following cepstral analysis calculates the 13 cepstral coffecients [5]. For the RASTA-PLP features, an additional filtering is applied after decomposition of the spectrum into critical bands. This RASTA filter suppresses the low modulation frequencies which are supposed to stem from channel effects rather than from speech characteristics. The PLP and RASTA-PLP streams were augmented by their delta features. For the extraction of the MSG features the frequency domain is divided into 1/4 octave bands, resulting in 14 bands, each of which is filtered with two modulation frequency pass-bands, the first ranging from 0-8 Hz, the second from 2-8 Hz. The two sets of 14 coefficients are then concatenated to give the MSG feature vector of 28 coefficients.

We work in the framework of HMM /MLP hybrid systems where the posterior probabilities at the output of the MLP are, after division by the priors, used as scaled likelihoods in the HMM for decoding [2,7]. The MLP uses 7 frames of context information (except for the MSG features where 9 frames are used) in order to better account for coarticulation effects and to model the phone changes in more detail. The hidden layer consists of 2,000 nodes (2770 for MSG), and the number of output nodes corresponds to the number of speech units in the digits and numbers part of the SpeechDat corpus.

We investigated the use of two different phone sets: (i) context-independent (CI) monophone models and (ii) context-dependent (CD) triphone models. The MLPs trained to estimate context-independent observation probabilities use 31 output nodes (one for silence), as only 30 monophones occur in the numbers part of the corpus. The remaining 7 nodes which usually correspond to the remaining monophones were not trained as these monophones did not occur in the digits and numbers part of the database. It is advantageous to model phonetic units with a sequence of probability distributions rather than with a single distribution only, in order to capture some of the dynamics of the phonetic segments. For this reason, the HMM state of each monophone model is repeated three to six times, depending on the respective monophone.

In order to better exploit the large acoustic input available to the MLP, context-dependent triphone models were investigated next. The use of triphones implies enlarging the output layer of the MLP. More (speech unit) classes at the output of the MLP renders the MLP more difficult to train and increases the need for more training data. For the digits and numbers task, this is still feasible as the number of occurring triphones is limited and the size of the MLP's output layer will not increase too much. To train the MLPs to output posterior probabilities for triphone models, we need a frame-level alignment for the triphones. For this, we substituted in the monophone-based alignment each monophone label by a new label which depended on both the monophone's left and right context: e.g. the monophone transcription of the word 'dois' (two) 'd of y ch' will result in the triphone transcription: '?-d-of d-of-y of-y-ch y-ch-?'. (The '?' marks the begin and end of a word.) This gave us a set of 151 triphone labels (word-internal only), used at the output of the context-dependent MLPs. This alignment was then used to train these context-dependent neural nets. The triphone HMM models use 3 states for duration modeling. Only the silence model uses just one state without duration modeling.

3.3 Vocabulary and Language Modeling

The vocabulary consists of 51 words for which an internal transcription was available. These words cover the 10 isolated digits, the 2 female forms "*uma*" and "*duas*", and the natural numbers as described in Sect. 2.1. Only 30 of the Portuguese phones actually occur in the digits and numbers, so that the phone set could be restricted to these 30 phones. The language model (LM) was set up on the training utterances, using the CMU-Cambridge Language Modeling Toolkit V2.05. The Good-Turing method was used to estimate the closed-vocabulary, back-off bigram LM which contains 2601 bigrams. Missing bigram combinations which did not occur in the training data were manually added. The perplexity of the LM on the test set is 10.73.

4 Experiments and Results

Experiments were carried out with HMM/MLP hybrid recognizers employing PLP [5], RASTA [6] or MSG [3] features. The results of the three systems are given in Table 4 for both the context-independent (CI) monophone models and the context-dependent (CD) triphone models.

Table 4. % Word error rates (WERs) of each of the three feature streams as employed in our HMM/MLP hybrid recognizer

	CI models	CD models
PLP	7.2	6.6
MSG	7.3	6.8
RASTA	8.0	8.4

The recognizers employing PLP or MSG features resulted in lowest word error rates (WERs) for both CI and CD modeling. The RASTA features which are also PLP-based but include a further filtering during feature extraction gave significantly worse results. RASTA filtering is usually necessary in noise corrupted speech and speech recorded over very different telephone lines. Although the SpeechDat database was collected over a large set of different telephone connections, the clean part of the SpeechDat corpus which we chose for these experiments seems to be rather homogeneous and does not need any additional noise filtering.

For the PLP and MSG feature sets, it is the context-dependent triphone modeling which enhances recognition performance, due to its modeling of larger contexts and coarticulation effects. In the case of the PLP features, the improvement in WER is significant. For the MSG features, the difference is not significant at a confidence level of 97.5%.

These results can e.g. be compared to results reported on the telephone-recorded OGI Numbers 1995 database [1] where authors report WERs of 6.8% using RASTA-PLP features and 9.8% using MSG features [9], and about 7.1% using PLP featuers [4].

5 Conclusions

In this article, we presented the Portuguese SpeechDat database, which is the first telephone-based, large-vocabulary speech database available in (European) Portuguese. We described its main features, such as recording conditions, speakers, vocabulary, and linguistic annotations, and the development of a speech recognizer trained and tested on the (clean) digits and numbers part of this corpus. The results show competitive performance to state-of-the-art (numbers) recognizers developed on telephone databases in other languages. We plan on extending our work on the Portuguese SpeechDat database to large-vocabulary recognition. Moreover, we want to use the noise annotations of the better labeled first part of the database in order to investigate and develop noise models which will help us to better annotate also the second part of the corpus. The final goal is to create speech recognizers robust to various kinds of different speaker, line and background noises.

Acknowledgments. Astrid Hagen was supported by the Portuguese FCT (Fundação para a Ciência e a Tecnologia) scholarship SFRH/BPD/6757/2001. Additionally, this work was partially funded by the FCT project POSI/33846/PLP/2000. INESC-ID Lisbon had support from the POSI Program of “Quadro Comunitário de Apoio III”.

References

1. Center for Spoken Language Understanding, Department of Computer Science and Engineering, Oregon Graduate Institute. Numbers Corpus, Release 1.0, 1995.

2. H. Bourlard and N. Morgan. *Connectionist Speech Recognition. A Hybrid Approach*. Kluwer Academic Publishers, 101 Philip Drive, Assinippi Park, Norwell, Massachusetts 02061 USA, 1994.
3. S. Greenberg and B.E.D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1647–1650, 1997.
4. Astrid Hagen. Robust speech recognition based on multi-stream processing. PhD thesis, Département d'informatique, École Polytechnique Fédérale de Lausanne, Switzerland, 2001.
5. H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.
6. H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. RASTA–PLP speech analysis technique. *IEEE Trans. on Signal Processing*, 1:121–124, 1992.
7. N. Morgan and H. Bourlard. Continuous speech recognition. *IEEE Trans. on Signal Processing*, pages 25–41, 1995.
8. SPEECHDAT. European speech databases for telephone applications (EU-project LRE-633140). In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1997.
9. S.L. Wu, B. Kingsbury, N. Morgan, and S. Greenberg. Incorporating information from syllable-length time scales into automatic speech recognition. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1:721–724, 1998.