Cypriot Speech Database: Data Collection and Greek to Cypriot Dialect Adaptation

Nikos Fakotakis

Wire Communications Laboratory
Electrical and Computer Engineering Dept.
University of Patras
265 00 Rion, Patras, Greece
fakotaki@wcl.ee.upatras.gr

Abstract

This paper describes the Cypriot Greek speech database collected in the framework of the European project OrienTel (IST-2000-28373) and the acoustic models adaptation techniques that were applied in order to perform dialect adaptation from Greek to Cypriot. Greek and Cypriot Greek share the same phoneme set. However, there are some differences in the way the same phonemes are pronounced. That is, Cypriot Greek may be considered as a variation of standard Greek. Utterances from 500 speakers are used (450 for training, that is, performing adaptation, and 50 as testing material). Two tools are available for training, adaptation and evaluation of the acoustic models. These are the Wire Communications Laboratory (WCL) recognition tool and the Hidden Markov Models toolkit (HTK). For both recognition engines Greek acoustic models were already available using the SpeechDat-II Greek telephone database. Two well-known techniques are applied for adapting the Greek acoustic models to the new data: Maximum Likelihood Linear Regression (MLLR) and Maximum A-Posteriori (MAP) adaptation. Pure Cypriot Greek models are also trained using only the Cypriot Greek database, to be compared with the adapted ones. Preliminary results show a small improvement in the performance of the adapted models over the pure Greek and Cypriot Greek models.

Introduction

The goal of this work is to describe the Cypriot Greek database collected in the framework of the European project OrienTel (IST-2000-28373, www.orientel.org) and the acoustic models adaptation techniques that were applied in order to perform dialect adaptation from Greek to Cypriot.

The main focus of OrienTel is the production of speech databases for the Mediterranean and Middle East countries, which will further enable the project's participants to design and develop multilingual speech-based applications. Therefore, the collected data is used for creating multilingual acoustic models and perform dialect or foreign accent adaptation, according to the region. For example, multilingual acoustic models and lexica are created for French, colloquial Arabic and standard Arabic in Morocco, Greek acoustic models are adapted to Cypriot Greek ones in Cyprus, etc.

Despite the fact that Cyprus had been under British occupancy for a long period, the native language of the Cypriots still remained the same (Cypriot Greek). English was additionally adopted by the habitants as a foreign language that would help in commercial and administrative purposes.

This paper focuses on the Cypriot database and Greek to Cypriot dialect adaptation. Greek and Cypriot Greek share the same phoneme set. However, there are some differences in the way the same phonemes are pronounced. That is, Cypriot Greek may be considered as a variation of standard Greek, such as the variations existing within the Greek territory (Greek as spoken in Crete, etc.) or the dialects spoken by Greek immigrants (United States, Australia, etc.).

As soon as the data collection is completed it will comprise 1000 recordings of Cypriot Greek and another 1000 recordings of English as spoken by Cypriots, all recorded through the fixed or mobile network. The experiments that are described in this paper are carried out with 500 recordings of Cypriot Greek, that is utterances taken from 500 Cypriot speakers.

Database Description

The database has been designed according to the OrienTel design specification based on the former SpeechDat projects (Höge et al., 1997).

Calls are recorded from the fixed or mobile telephone network via an ISDN line connection. The signals are stored directly in digital format using A-law coding. They are recorded with a sampling rate of 8 kHz, 8-bit quantization with the least significant byte first ("lohi" or Intel format) as (signed) integers. Every speech file is accompanied by the corresponding SAM label file, which contains information about the recording conditions, the speaker, the transcription of the utterance, etc.

The database includes utterances for isolated digits, digit strings, application words, application word phrases, dates, times, directory assistance names, phonetically rich words, phonetically rich sentences and spontaneous speech (Gedge et al., 2002). Care has been taken to adequately cover all phonemes. The speakers recruited vary in age and the sessions have been recorded in various environments. Each speaker utters 49 different sentences. The orthographic and phonetic transcriptions of the spoken utterances also follow the SpeechDat conventions (Van den Heuvel et al., 2001). The recordings are annotated using a graphical parametric language-independent tool (Georgila et al., 2000a), which was first developed for the annotation of SpeechDat and SpeechDat-Car, and then it was extended to incorporate the additional features of OrienTel.

For the OrienTel recordings, Cyprus has been divided into two dialect areas. The North-East part of the island (Nicosia region) uses the most common dialect (spoken by 70% of the population) and the South-West region

(Paphos) uses the Paphos dialect that differs significantly from the common dialect.

Data Used for Adaptation

The dialect adaptation between Greek and Cypriot Greek is divided in two phases. This stepwise approach aims at defining the exact methodology that will be used for adaptation and spotting possible problems or difficulties in the first stage, so that the second phase of adaptation is performed in the best possible way. In the first phase utterances from 500 speakers are used (450 for training, that is, performing adaptation, and 50 as testing material) whereas in the second stage the whole database is used, that is, 900 speakers for training and 100 for tests.

More specifically, as training material the total number of available utterances is 49 x 450 = 22050 in the first phase of adaptation and 49 x 900 = 44100 in the second stage. However, in this paper only the first phase will be described since data is not available yet to proceed with the second stage.

Each speaker session contains 50 files. However, item codes D2 and D4 correspond to prompted date phrases for the Western and the Islamic calendar respectively. Thus, since the Islamic calendar is not used, items D2 and D4 are exactly the same and therefore each speaker session includes 49 different items. Only the utterances that contain mispronunciations, illegible parts and truncations are excluded.

As testing material 5 different test sets are selected. These sets are extensively described in (Gedge et al., 2002; Georgila ., 2003).

- *Isolated digits*
  I1: single isolated digit
  B1: sequence o f 10 isolated digits

- *Digit strings*
  C1: prompt sheet number (6 digits)
  C2: telephone number (8-13 digits)
  C3: spontaneous telephone number
  C4: credit-card-like number (14-16 digits)
  C5: PIN code (6 digits)

- *Application words*
  A1-A6

- *Dates*
  D1: birth date (spontaneous)
  D2: prompted date phrase (Western calendar).
  D3: relative and general date expression

- *Directory assistance names*
  O1: personal first name (spontaneous)
  O2: city of childhood (spontaneous)
  O3: most frequent cities (both local and foreign)
  O5: most frequent companies/agencies
  O7: personal name (first name and family name)

After discarding the utterances that contain mispronunciations, illegible parts and truncations, the final number of utterances used for training and testing in the first phase of adaptation is described in Table 1.

Number of utterances used					
Training	Test				
All items	*Isolated digits*	*Digit strings*	*Appl. words*	*Dates*	*Directory assistance names*
21935	100	228	300	141	233

Table 1: Number of utterances used as training and testing speech material

Speech Recognition Systems

Two recognisers are available. The first one has been developed by WCL and is based on a set of 808 basic units (phonemes and syllables i.e. two-phone, three-phone, four-phone and five-phone combinations of two or more consonants always ending in a vowel), described by a 5-state left to right continuous HMM. During the on-line recognition procedure the Frame Synchronous Viterbi Beam Search algorithm is used in order to produce the most probable sequences of basic recognition units taking into consideration the transition probabilities between them. Then the 10-best word sequences (hypotheses) are formed using the current lexicons and bigram probabilities between the words.

The second speech recogniser is built with the HTK Hidden Markov Models toolkit (Young et al., 2002). Each one of the 37 Greek monophones is described by a 5-state left to right continuous HMM, the parameters of which have been defined during the off-line embedded Baum-Welch re-estimation procedure. This set of monophone HMMs is used to create context-dependent triphone HMMs. This is done in two steps. Firstly, the monophone transcriptions are converted to triphone transcriptions and a set of triphone models are created by cloning all the monophones and re-estimating, which leads to a very large set of models, that is 6139, and relatively little training data for each model. Secondly, similar acoustic states of these triphones are tied to ensure that all state distributions can be robustly estimated. Lattices are used for language modelling.

In order to train both recognisers we used the SpeechDat-II Greek telephone database (Georgila, 2000b; Chatzi et al., 1997). This database is a collection of Greek annotated speech data from 5000 speakers (each individual having a 12-minute session). We made use of utterances taken from 3000 speakers in order to train our system. Each input speech signal waveform is sampled at 8 kHz, pre-emphasised by the filter $H(z)=1-0.97z^{-1}$ and subsequently windowed into frames of 20 ms duration at a frame rate of 10 ms using a Hamming window. The features that are extracted are Mel Frequency Cepstral Coefficients with their temporal regression coefficients of first and second order.

The phoneme set used is the standard SAMPA for Greek, which applies also to Cypriot Greek (SAM-PA).

Adaptation Techniques

In order to build robust acoustic models for Cypriot Greek a lot of data must be available. Considering that only 500 speakers are available in the first phase of adaptation and 1000 speakers in the complete database, adaptation techniques are employed so that the Greek acoustic models are adapted in order to cover variations due to the Cypriot Greek dialect.

Adaptation techniques can be used in various different modes. If the true transcription of the adaptation data is known then it is termed supervised adaptation, whereas if the adaptation data is unlabelled then it is termed unsupervised adaptation. In the case where all the adaptation data is available in one block, then this is termed static adaptation. Alternatively adaptation can proceed incrementally as adaptation data becomes available, and this is termed incremental adaptation.

Two well-known techniques for adapting acoustic models to new data are maximum likelihood linear regression (MLLR) and maximum a-posteriori (MAP) adaptation. In order to adapt the acoustic models of the WCL recogniser MLLR is used, whereas both MLLR and MAP are applied for adapting the HTK-based acoustic models using the HEAdapt tool. HEAdapt performs offline supervised adaptation using maximum likelihood linear regression (MLLR) and/or maximum a-posteriori (MAP) adaptation, while unsupervised adaptation is supported by HVite (using only MLLR). In our case, supervised adaptation is performed because the transcriptions of the adaptation data are known. Currently, MLLR adaptation in HTK can be applied in both incremental and static modes while MAP supports only static adaptation. For our experiments static adaptation is selected because all data is available from the beginning.

Maximum Likelihood Linear Regression (MLLR)

Maximum likelihood linear regression or MLLR computes a set of transformations that will reduce the mismatch between an initial model set and the adaptation data. More specifically MLLR is a model adaptation technique that estimates a set of linear transformations for the mean and variance parameters of a Gaussian mixture HMM system. The effect of these transformations is to shift the component means and alter the variances in the initial system so that each state in the HMM system is more likely to generate the adaptation data.

The transformation matrix used to give a new estimate of the adapted mean is given by

$$\boldsymbol{\mu} = \boldsymbol{W} \boldsymbol{x}$$

where $\boldsymbol{W}$ is the n × (n + 1) transformation matrix (where n is the dimensionality of the data) and $\boldsymbol{x}$ is the extended mean vector,

$$\boldsymbol{x} = [w \ \mu_1 \ \mu_2 \dots \mu_n]^T$$

where w represents a bias offset.

Hence $\boldsymbol{W}$ can be decomposed into

$$\boldsymbol{W} = [\ \boldsymbol{b}\,\boldsymbol{A}\ ]$$

where $\boldsymbol{A}$ represents an n × n transformation matrix and $\boldsymbol{b}$ represents a bias vector.

The transformation matrix $\boldsymbol{W}$ is obtained by solving a maximisation problem using the Expectation-Maximisation (EM) technique. This technique is also used to compute the variance transformation matrix (Young et al., 2002).

Maximum A-Posteriori adaptation (MAP)

This adaptation process is sometimes referred to as Bayesian adaptation. MAP adaptation involves the use of prior knowledge about the model parameter distribution. Hence, if we know what the parameters of the model are likely to be (before observing any adaptation data) using the prior knowledge, we might well be able to make good use of the limited adaptation data, to obtain a decent MAP estimate. This type of prior is often termed an informative prior. Note that if the prior distribution indicates no preference as to what the model parameters are likely to be (a non-informative prior), then the MAP estimate obtained will be identical to that obtained using a maximum likelihood approach.

For MAP adaptation purposes, the informative priors that are generally used are the dialect independent model parameters. For mathematical tractability conjugate priors are used, which results in a simple adaptation formula.

If the likelihood of the adaptation data is small, then the mean MAP estimate will remain close to the dialect independent component mean. With MAP adaptation, every single mean component in the system is updated with a MAP estimate, based on the prior mean, the weighting and the adaptation data. Hence, MAP adaptation requires a new “dialect-dependent” model set to be saved.

One obvious drawback to MAP adaptation is that it requires more adaptation data to be effective when compared to MLLR, because MAP adaptation is specifically defined at the component level. When larger amounts of adaptation training data become available, MAP begins to perform better than MLLR, due to this detailed update of each component (rather than the pooled Gaussian transformation approach of MLLR). In fact the two adaptation processes can be combined to improve performance still further, by using the MLLR transformed means as the priors for MAP adaptation. In this case components that have a low occupation likelihood in the adaptation data, (and hence would not change much using MAP alone) have been adapted using a regression class transform in MLLR (Young et al., 2002).

Evaluation

The following recognition systems in both adaptation phases were trained and evaluated:

WCL recogniser
Greek models
Cypriot Greek models
adapted models (MLLR)

HTK recogniser
Greek models
Cypriot Greek models
adapted models (MLLR)
adapted models (MAP)
adapted models (MLLR and MAP)

The results of the tests carried out with the WCL recogniser are given in Table 2 whereas Table 3 shows the results of the tests conducted with HTK. However, these are preliminary results and experiments are still being carried out.

In tables 2 and 3, we can see that there is a small improvement in the performance of the adapted models over the pure Greek and Cypriot Greek models and this is more obvious for the HTK recogniser. The poor performance of the pure Cypriot Greek models is due to the inadequate training data (only 450 speakers).

WCL recogniser – Word accuracy (%)					
Models type	*Isolated digits*	*Digit Strings*	*Appl. Words*	*Dates*	*Directory assistance names*
Greek models	85.5	56.2	77.3	61.1	52.4
Cypriot Greek models	70.4	46.8	69.5	49.8	45.6
Adapted models (MLLR)	86.1	55.0	79.2	62.0	54.0

Table 1: Word accuracy reached with the WCL recogniser and different model types and testing material

HTK recogniser – Word accuracy (%)					
Models type	*Isolated digits*	*Digit Strings*	*Appl. words*	*Dates*	*Directory assistance names*
Greek models	90.1	62.0	85.1	68.3	65.2
Cypriot Greek models	85.4	51.7	67.9	52.5	56.3
Adapted models (MLLR)	91.0	64.3	85.8	69.0	65.0
Adapted models (MAP)	90.8	62.8	86.0	67.6	64.5
Adapted models (MLLR and MAP)	91.4	64.1	86.3	69.2	65.1

Table 2: Word accuracy reached with the HTK recogniser and different model types and testing material

Conclusions

In this work, the Cypriot Greek speech database collected in the framework of the European project OrienTel (IST-2000-28373) is described. Moreover, adaptation of the Greek acoustic models to the Cypriot Greek dialect is performed. Two different recognition systems are used and two different adaptation methods are applied. Greek acoustic phonemes were already available using the SpeechDat-II Greek telephone database. Experiments are carried out for both recognisers (WCL and HTK) and adaptation techniques (MLLR and MAP) as well as for 5 different test sets with utterances taken from 500 speakers (450 for training, that is, performing adaptation, and 50 as testing material). Pure Cypriot Greek models are also trained using only the Cypriot Greek database to be compared with the adapted ones. Preliminary results show a small improvement in the performance of the adapted models over the pure Greek and Cypriot Greek models.

References

Chatzi, I., Fakotakis, N., Kokkinakis, G. (1997). Greek speech database for creation of voice driven teleservice. EUROSPEECH, Vol. 4, pp. 1755-1758, Rhodes, Greece.

Gedge, O., Shammass, S., Moreno, A., Choukri, K., Emam, O., Zitouni, I., Heuft, B. (2002). Speech Database Design. Deliverable D2.1 of the Orientel Project.

Georgila, K., Fakotakis, N., Kokkinakis, G. (2000a). A Graphical Parametric Language-Independent Tool for the Annotation of Speech Corpora. LREC, Vol. 3, pp. 1537-1542, Athens, Greece.

Georgila, K. (2000b). Greek SpeechDat-II database. Documentation file included in the database CD-ROMs.

Georgila, K. (2003). Language-Specific Peculiarities for Cypriot Greek and English as Spoken by Native Speakers in Cyprus. Deliverable D2.3 of the Orientel project.

Höge, H., Tropf, H., Winski, R., Van den Heuvel, H., Haeb-Umbach, R., Choukri, K (1997). European speech databases for telephone applications. ICASSP, Vol. III, pp. 1771-1774, Munich, Germany.

SAM-PA, *Standards, Assessment, and Methods: Phonetic Aplhabets*. http://phon.acl.ac.uk/home/sampa/home.htm

Van den Heuvel, H., Moreno, A., Omologo, M., Richard, G., Sanders, E. (2001). Annotation in the SpeechDat Projects. International Journal of Speech Technology, 4(2), pp. 127-143.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. (2002). The HTK Book (for HTK Version 3.2).