Language-identification based on Cross-Language Acoustic models and Optimised Information Combination

Ove Andersen and Paul Dalsgaard
Center for PersonKommunikation (CPK)
Aalborg University, Denmark

ABSTRACT

This work is concerned with the subject of language-identification (LID). Two central issues are addressed. The first is to analyse the trade-off between detailed acoustic modelling and robust estimation of acoustic and language models. The second to find the optimal combination of acoustic and language scores for language-identification.

Experiments are carried out using the three languages American-English, German and Spanish from the OGI-TS database. It is shown that on the average the acoustic modelling is able to recognise 46.3% of the phones correctly across the three languages. Insertion and deletion rate is 35.7% and 6.6%, respectively. Language-identification performance is 82.6% with the full set of acoustic models. The performance is increased to 83.7% after having conducted 80 iterations of a hierarchical clustering in which phones are merged across the languages.

1. INTRODUCTION

Definitions of individual phonemes within and across a number of languages - for instance as given in [1]- clearly demonstrate that some phonemes are very similar as seen from their articulatory representation. The research presented in this paper is focussed on analysing and verifying the possible merging of speech segments within and across a number of languages.

A data-driven technique is established for identifying those phones which are found similar enough to be merged into one acoustic model. A number of merged phones are employed together with the remaining non-combined phones in two experiments. The first is a phone recognition experiment, the second is a LID-experiment encompassing a number of tests. In each of the LID-tests either a linear or a non-linear transformation is superimposed on the outputs from the acoustic and the language models. The aim of these tests is to identify the transformation which is best suited for optimal language-classification.

2. ARCHITECTURE

The LID-system - shown in Figure 1 - is used to test language-identification among three languages which are all taken from the spontaneous telephone speech corpus OGI_TS [2]. The languages are American-English (US), German (GE) and Spanish (ES). The LID-system is composed of a common acoustic signal preprocessor and two modules each contributing to the processing involved in the task of language-identification. The first of these modules performs the phone and language decoding, the second transforms the parameters from the decoding module and classifies the language.

The common acoustic signal preprocessor calculates 12 RASTA filtered MFCC's, their first derivatives and the delta-log-energy. The phone and language decoding module consists of three parallel branches. In each of these the phone recogniser matches the acoustic parameters to the acoustic models used by that recogniser. The output from each recogniser is further matched against three language models.

The combined output **X** from all language models and from all recognisers are used as input to the 'information combination and the language-classification' module (ICLC). This module enforces a transformation onto the parameters **X** and estimates the most probable language given the acoustic input.

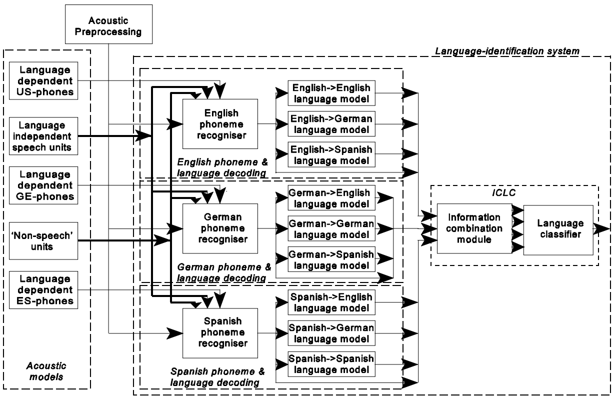
Acoustic Preprocessing
Language-identification system
Language dependent US-phones
Language independent speech units
Language dependent GE-phones
'Non-speech' units
Language dependent ES-phones
Acoustic models
English phoneme recogniser
English phoneme & language decoding
English->English language model
English->German language model
English->Spanish language model
German phoneme recogniser
German phoneme & language decoding
German->English language model
German->German language model
German->Spanish language model
Spanish phoneme recogniser
Spanish phoneme & language decoding
Spanish->English language model
Spanish->German language model
Spanish->Spanish language model
ICLC
Information combination module
Language classifier

Figure 1. Architecture of language-identification system

3. PHONE SIMILARITIES

The research is based on the postulate that the initial set of N *language-specific* phones φ_k across a number of languages, e.g. for the three languages chosen for this work:

$$\phi = \phi_{US} \cup \phi_{GE} \cup \phi_{ES} = \{\varphi_1 \ \varphi_2 \cdot \cdot \ \varphi_N\}$$

can be separated into a common group Φ_{li} of *language-independent acoustic models* and three groups of remaining *language-dependent* phone models $\Phi_{ld,\ US}$, $\Phi_{ld,\ GE}$ and $\Phi_{ld,\ ES}$ from the involved training languages.

The total initial set of phones is N = 113 (40 for US, 41 for GE and 32 for ES) is used for the three training languages.

Based on the training material, each of the initial N phones is modelled by a hidden Markov model (details in section 4). The models are used to initialise a data-driven iterative procedure the aim of which is to identify a group of combinable language-independent acoustic models and three groups of non-combinable language-dependent phone models. A language-independent acoustic model is established by merging the two most similar phones. These may be drawn from the group of *language-specific phones* or the group of *language-independent acoustic models*. The similarity is measured by using 200 randomly selected speech segments for each acoustic model. All samples are used as input to all models whereby average log-likelihoods for a speech segment given an acoustic model are established. The likelihoods constitute the basis for calculating the similarity. Further details are given in [3].

It is emphasized that the data-driven, iterative process for the selection of the language-independent acoustic models in a flexible way makes it possible to test the dependency of the scoring of the language-identification system upon the number of preselected language-independent acoustic models.

4. PHONE MODELS

Each of the *language-independent acoustic models* and the *language-dependent phone models* is modelled by a left-to-right, context independent, three state CDHMM including one skip and having two mixtures in each state.

In addition to these speech segment models, all non-speech segments contained within the training corpus are represented by a group of *non-speech models*, each of which is modelled by an ergodic four state CDHMM with two mixtures in each state.

5. LANGUAGE MODELS

The output from each of the phone recognisers is used to train three language models. Each language models represents bigram probabilities, which are calculated on the basis of the decoded output from the recogniser from which they receive their input. The language model identified by *English->German* in Figure 1, is trained on *German* spontaneous speech. The acoustic parameters from German speech signals are submitted to a phone decoding performed by the *English* phone recogniser.

The back-off bigram is given by:

$$p(i,j) = \{ \begin{array}{ll} (N(i,j) - D)/N(i) & \text{if } N(i,j) > t \\ b(i)p(j) & \text{otherwise} \end{array}$$

where N(i,j) is the number of transitions from segment label *i* to segment label *j*. The total number of occurrences for label *i* is given by $N(i) = \Sigma\, N(i,j)$, $1 \le j \le Q$, where Q is the number of acoustic models. D is a discount which has the effect of smoothing the bigram probabilities. The back-off weight b(i) ensures that $\Sigma\, p(i,j)=1$, $1 \le j \le Q$.

The uni-gram probability p(j) is given by:

$$p(j) = \{ \begin{array}{ll} N(j)/N & \text{if } N(j) > u \\ u/N & \text{otherwise} \end{array}$$

where u is the uni-gram floor and $N = \Sigma \max [N(j),u]$, $1 \le j \le Q$.

6. LANGUAGE CLASSIFICATION

The ICLC-module takes as input the combined parameters **X** from the phone and language decoders on which it performs a mapping into the most likely language class.

Several linear and non-linear techniques have been evaluated, namely : 1) Linear Discriminant Analysis (LDA), 2) k-Nearest Neighbour Rule (k-NNR), 3) Mahalanobis distance measure (MDM) and 4) Oblique Classification Tree (OCT). The processing involved with each of these techniques is briefly introduced below.

6.1. LDA transformation

A linear discriminant analysis [4] is introduced with the aim of maximal discrimination between the corresponding output classes. The result of the analysis is a transformation **W** which establishes a set of new parameters **Y** - possibly of reduced size - which are used by the language classifier to perform an optimal discrimination between the a-priori defined classes. The result is the projection:

$$\boldsymbol{Y} = \boldsymbol{W}^T \boldsymbol{X}$$

which maximises the function:

$$J(\boldsymbol{W}) = \frac{\boldsymbol{W}^T \boldsymbol{S}_B \boldsymbol{W}}{\boldsymbol{W}^T \boldsymbol{S}_W \boldsymbol{W}}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ are the between and within class scatter matrices, respectively. The solution is:

$$\boldsymbol{W} = \boldsymbol{S}_W^{-1}(\boldsymbol{m}_i - \boldsymbol{m}_j)$$

where $\mathbf{m}_i$ and $\mathbf{m}_j$ are means of the classes *i* and *j*.

6.2. k-NNR transformation

The k-Nearest Neighbour Rule is a classification scheme which stores training vectors together with their associated class label in a code book. During classification of an unknown vector **X** the distance between this vector and each of the code book entries is measured. The label associated with the nearest training vector is used as the estimated class membership for the test vector. This corresponds to 1-NNR. For k-NNR the *k* nearest code book entries are found and a decision is taken according to a voting procedure. It is important that ties can be avoided or resolved. Hence, for two-class cases k is normally an odd number.

In the present three-class problem ties are resolved by selecting the class with minimum accumulated distance.

6.3. MDM transformation

A very common and very simple metric is the Mahalanobis distance measure (MDM). Compared to the Euclidian metric it has the advantage of removing biases and taking the correlation between the elements of the feature vector **X** into account.

For each of the *c* classes *i* a mean vector $\mathbf{m}_i$ and a corresponding covariance matrix $\mathbf{C}_i$ is estimated. An unknown feature vector **X** is simply classified by measuring the Mahalanobis distance from **X** to each of the classes and assigning **X** to the class for which the distance is minimum.

6.4. OCT transformation

A classification tree consists of nodes and branches. Each node represents a single test or decision. The test is normally binary and - depending on whether the result is true or false - the tree branches either left or right to the new node. Descending down the tree a terminal node is reached and a class decision is taken.

The test at each node may have the form x_i>k where x_i is an element of the input vector and k is a constant. This divides the input space with axis-parallel hyper-planes and the trees are called *axis-parallel decision trees.*

Alternatively, the tests at each node can be formed by a linear combination of inputs:

$$\sum_{i=1}^{d} a_i x_i + a_{d+1} > 0$$

where x_i and a_i are inputs and real-valued coefficients, respectively. The dimension of the input vector **X** is d. These tests represent hyper-planes with an oblique orientation to the axes and are consequently called *oblique decision trees* (OCT) [5].

It is normally necessary to prune decision trees to eliminate problems with overtrained trees. A technique called Cost Complexity Pruning is applied, see [6] for details. The idea is to reserve a part (e.g. 10%) of the training data for pruning. A set of decreasing sized sub-trees of the original tree is found. These are then used to classify the pruning data. The smallest tree with highest performance is selected.

7. DATABASE

The database used is the OGI-TS spontaneous telephone speech database [2]. The three European languages American-English, German and Spanish are selected for the tests performed within this work.

The available 'before-tone' data for each language are divided into two sets of approximately equal size; one for training and one for testing. Thus each of the files used for testing has a duration of up to 50 secs.

8. EXPERIMENTS

Two experiments have been conducted. The first is a phone recognition experiment in which a varying number P of merged acoustic models are used by the recognisers.

The second is a number of language-identification tests in which the transformation which gives optimal LID-score is firstly identified, and secondly detailed LID-analyses are conducted with the selected transformation and in testing the LID-score for a varying number P of acoustic models used by the recognisers.

8.1. Phone recognition experiment

Table I shows the results of a number of phone recognition experiments. In interpreting the results it is emphasized that the test results are given for telephone quality test data, i.e. collected over telephone channels of varying quality and in using a large number of different hand-sets.

Table I. Phone recognition using a varying number P of *language-independent acoustic models,* ten '*non-speech*' models together with the non-combined *language-dependent phone models* for each of the languages

No. P of merged acoustic models	% US	% GE	% ES	% Average
0	43.0	44.6	51.4	46.3
10	44.7	41.8	43.7	43.4
20	39.9	42.0	45.9	42.6
30	43.8	42.6	43.9	43.1
40	44.4	43.5	44.7	44.2
50	44.0	43.1	43.5	43.5
60	45.7	44.6	32.6	41.0
70	48.1	44.6	34.4	42.3
80	47.0	43.7	35.9	42.2
90	49.5	46.9	39.1	45.2

The test with P=0 corresponds to tests with *language-specific phone* models only. In the tests all models have been re-estimated twice. It is observed that the average phone recognition accuracy - as seen from an overall point of view - stays almost constant.

8.2. Language-identification experiment

8.2.1. Linear/non-linear transformations in LID

Table II shows the results for a number of language-identification tests using a number of linear and non-linear techniques for 'information combination'. The length of the input **X** vector is twelve, i.e. all parameters from the decoding module are utilised. The experiments are for P=0, i.e. only *language-specific acoustic models* are used.

The results show the highest average LID-score for the 5-NNR transformation. This transformation is applied in the following more detailed tests.

8.2.2. 5-NNR transformation and varying P

The results are shown in Table III. The tests are conducted with a varying number P of language-independent acoustic models and using all information available from the decoding modules, i.e. length [**X**]=12.

Table II. Results of LID-experiments using language-specific acoustic models in the phone recognisers and all outputs from phone and language decoders are used as input to the ICLC-module.

Type of mapping technique	% LID- scores for P = 0 and length [X] = 12			
	US	GE	ES	Average
LDA	98.6	60.0	64.2	76.7
1-NNR	84.1	70.0	81.1	79.1
5-NNR	97.1	68.0	77.4	82.6
MDM	98.6	46.0	64.1	72.7
OCT	84.5	62.2	80.0	76.6

Table III. Results obtained using the 5-NNR transformation, calculated on the basis of all information from the phone and language decoders and for a varying number P of merged acoustic models.

No. P of merged acoustic models	% LID- scores for varying P and length [X] = 12			
	US	GE	ES	Average
0	97.1	68.0	77.4	82.6
10	94.2	58.0	69.8	76.2
20	94.2	60.0	79.3	79.7
30	91.3	48.0	81.1	75.6
40	88.4	50.0	71.7	72.1
50	89.9	58.0	60.4	71.5
60	88.4	58.0	66.0	72.7
70	91.3	72.0	71.7	79.7
80	98.6	72.0	75.5	83.7
90	89.9	56.0	66.0	72.7

It is observed that a relative high average language-identification score is maintained even for a large number of language-independent acoustic models used by the recognisers. The highest performance is observed after the merging of 80 acoustic models.

8.2.3. 5-NNR transformation, P=30, varying length [X]

Figure 2 show the results of a number of LID-tests conducted using the 5-NNR transformation for a varying P and selected parameter sets used as input to the ICLC-module. The results given by the graph '3 languages' correspond to input from the language models *English->English, German->German* and *Spanish->Spanish* only.

The results show on the one hand that the importance of the acoustic models decreases for increasing values of P.

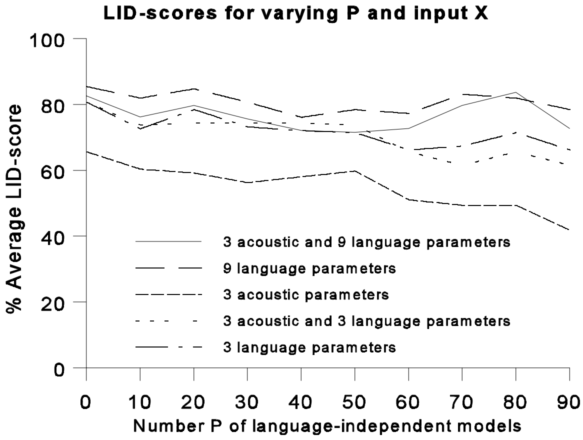

Figure 2. LID-results for different input **X** and varying P

On the other hand that the language models seem not to be too sensitive to varying values of P.

9. CONCLUSIONS

The original contribution of the approach being presented in this work is twofold:

- it is demonstrated that some of the acoustic models can be trained on data from more than a single language without any negative effects on the performance of the language-identification system.

The experiments showed that the highest performance is achieved after the merging of about 80 acoustic models.

- it is verified that the selection of a suitable scheme for combining scores from acoustic and language modelling is critical. The difference in performance between the tested schemes is up to 6%.

10. ACKNOWLEDGEMENTS

The work has been conducted on the basis of funding given to CPK from the Danish Technical Research Council.

11. REFERENCES

[1] J L Hieronymus . 'ASCII Phonetic Symbols for the World's languages: Worldbet'. Journal of the International Phonetic Association, 1993.

[2] Y K Muthusamy, R A Cole, B T Oshika. 'The OGI multi-language telephone speech corpus', Proc. Int. Conf. ICSLP92, 1992, pp 895-898.

[3] P Dalsgaard and O Andersen, 'On the identification and use of Language-independent speech units in Language-identification'. Proceedings of CRIM/ FORWISS Workshop, Montreal, Oct. 7-8, 1996.

[4] D E Morrison. 'Multivariate Statistical Methods', McGraw-Hill Series in Probability and Statistics, Third Edition, 1990.

[5] S K Murthy, S Kasif and S Salzberg. 'A System for Induction of Oblique Decision Trees', Journal of Artificial Intelligence Research, pp 1-32, 1994.

[6] L Breiman, J H Friedman, R A Olshen and C J Stone. 'Classification and Regression Trees', Chapman&Hall, New York, 1984.