

LANGUAGE-IDENTIFICATION USING LANGUAGE-DEPENDENT PHONEMES AND LANGUAGE-INDEPENDENT SPEECH UNITS

*Paul Dalsgaard, Ove Andersen, Hanne Hesselager, Bojan Petek*¹
Center for PersonKommunikation, Aalborg University, Denmark

ABSTRACT

This paper reports on results from ongoing research on language-identification (LID) performed on the three languages: American-English, German and Spanish. The speech material used is from the Oregon Graduate Institute Spontaneous Telephone Speech Corpus, OGI_TS.

The baseline LID-system consists of three parallel phoneme recognisers each of which are followed by three language modelling modules each characterising the bigram probabilities.

The phoneme models used are derived on the basis of the combined speech corpus comprising the three languages. The phonemes are handled differently in analysis performed in two experiments. In the first experiment they are trained and tested language-specifically. In the second, they are separated into a number of groups, one of which contains those language-independent speech units which are similar enough to be equated across the training languages, the remaining containing the non-combinable language-dependent phonemes for each of the languages. A data-driven technique has been devised to separate the speech sounds contained within the training corpus into these groups. In order to prepare for an optimal separation between the input classes, a linear discriminant analysis is performed on the training speech material.

Results from a number of experiments show that average language-identification scores of close to 90% can be retained by the LID-system presented here even for a high number of language-independent speech units.

1. INTRODUCTION

This paper is based on similar LID-system approaches to the ones presented in [1] and [2], but two essential differences are being emphasised in the work presented here. The first is that of separating the speech sound models used by the system into a common group of *language-independent speech sounds*, and a separate group of remaining *language-dependent phonemes* for each of the training languages. The second is that the final language identification is based on an optimal linear classification based on a set of decision parameters which results from a linear discriminant transformation.

2. THE BASELINE SYSTEM

The baseline LID-system is shown in Figure 1. The speech signal parametrisation is performed by the common acoustic preprocessing module. Details are given in section 3.

The language identification part of the system consists of four modules. The first - the phoneme decoding module - consists of a set of parallel phoneme recognisers each of which performs the acoustic decoding on the basis of groups of acoustic models which are selected and tuned to a specific training language.

The output from each of the recognisers is fed into the second module - the language decoding module - which consists of a set of parallel language modules. Each language module consists of (three) language models each representing the bigram language model for a specific language. Each of the language models is trained on the decoded output from the corresponding phoneme recogniser given acoustic speech training material from that language.

Each language module in Figure 1 consists of three language models. Using the technique presented in this paper, however, this number can be chosen freely such that the language-identification system can handle more languages than the three shown given that the proper training corpora are available. The output parameters from each of the language modules and from each of the phoneme recognisers are all fed into a third module in which the parameters are submitted to a linear discriminant transformation. The transformed parameters are used in the fourth module - the classifier - which selects the most likely candidate language given the acoustic input.

3. PREPROCESSING

Three basic preprocessing methodologies have been tested in the experiments in order to analyse their potential in the specific task of phoneme recognition and the overall task of language-identification. It is emphasised that the selected methodology is likely to be dependent on the corpora used for training, and that the work presented in the paper deals with noise-contaminated spontaneous telephone speech, which is collected using a large number of different telephone hand sets. The three preprocessing schemes are:

- ▶ the standard Mel Frequency Cepstral Coefficients (12 MFCC, 12 Δ -MFCC and Δ -Log E) [3,4],
- ▶ the Forward-Backward DYNAMIC (FBDYN) Cepstrum vector. The dynamic cepstrum simulates the time-frequency characteristics of the forward and backward auditory masking in order to enhance the dynamic features of speech while reducing sensitivity to different channel characteristics. The Dynamic cepstrum is derived from the cepstrum vector using a two-dimensional time-frequency masking model based on results obtained from perception experiments [5], and

¹ Visiting researcher from University of Ljubljana, Slovenia and postdoctoral scholar of the Slovenian Science Foundation.

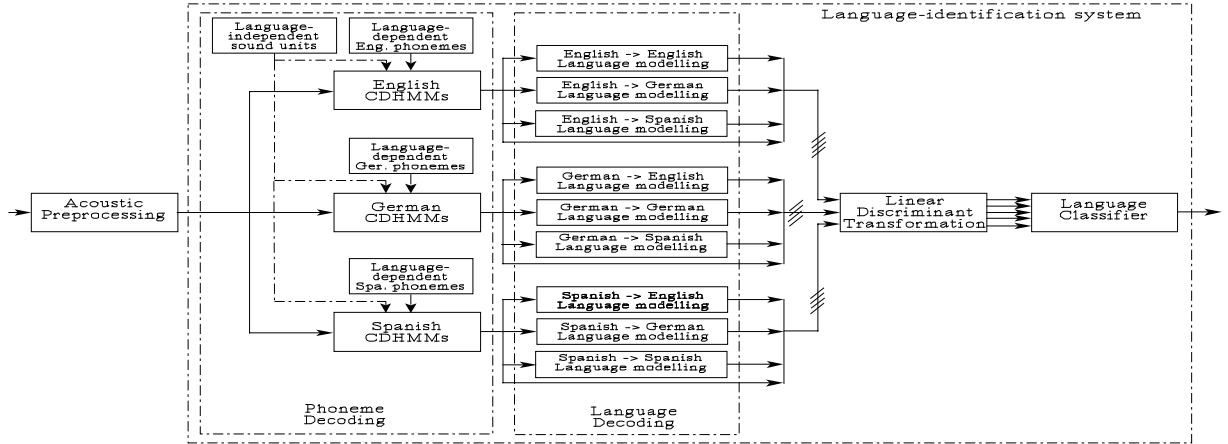


Figure 1. The baseline LID system.

► the RelAtive SpecTrAl feature vector resulting from the MFCC vector after being filtered by the liftering function [4].

4. TRAINING AND TEST CORPUS

The speech corpus is the Oregon Graduate Institute Spontaneous Telephone Speech Corpus, OGI_TS from which training and test material for the three languages American-English, German and Spanish is employed.

The phonemes are given by the training material which is annotated to the segmental level using the Wordbet set of symbols. The details are at the phonemic level as all diacritics are removed before the training is initiated. Those phonemes which have a representation of less than 45 realisations are merged with their closest equivalent symbol. Ten 'non-speech' segment types are used to model 'non-speech' events in the combined training corpus. For details, see [6]. Approximately 64 minutes of speech is used for the training, and 57 minutes for testing.

5. PHONEME GROUPS

The results of the experiments conducted in connection with this paper are based on two different ways of 'handling' the phonemes contained within the three training languages.

The first experiment is conducted on the basis of the set Φ (the total phoneme inventory) of N *language-specific* phonemes ϕ_k across the three languages. These phonemes are those from the OGI_TS corpus, but where all diacritics have been removed.

The second experiment is based on the separation of the set Φ across the three languages, i.e.

$$\Phi = \Phi_{US} \cup \Phi_{GE} \cup \Phi_{ES} = \{\phi_1, \phi_2, \dots, \phi_k, \dots, \phi_N\}$$

into a set Φ_{li} of *language-independent* (li) sound units and three set of remaining *language-dependent* (ld) phonemes $\Phi_{ld, US}$, $\Phi_{ld, GE}$ and $\Phi_{ld, ES}$. In previous work [7,8] we have employed two methodologies by which the total set of language-specific phonemes Φ can be separated.

5.1. Data-driven definition of language-independent speech units

The data-driven strategy for the merging of some of the phonemes across the three training languages into their equivalent speech units is presented here. The terminology 'speech unit' is used henceforth rather than phoneme as the iterative selection process normally lead to the merging of a number of phonemes across the training languages. In other words the data-driven *language-independent speech units* are those which emerge from the clustering speech sounds which are similar enough to be equated, but where each of them is a phoneme.

Based on the training material, each of the initial $N = 113$ language-specific phonemes - being trained on its corresponding acoustic data - is modelled by a hidden Markov model. The language-specific models are used to initialise an iterative procedure the results of which are the group of data-driven language-independent sound units and the three groups of language-dependent phonemes.

During each iteration, the strategy is to select the two most similar speech units and the methodologies used for measuring the similarities and for selecting the language-independent sound units are outlined below. The similarity of a phoneme and/or a speech unit is based on the calculation of measured log-probability $c(\phi_i, \phi_j)$ among all the phonemes or speech units ϕ_i and ϕ_j within the combined training corpus. The phoneme recogniser, which is subsequently used within the language-identification experiments, is used to establish the average per-frame values of the log-probabilities.

The iterative process as such is initialised on the basis of the language-specific set Φ of $N_0 = N$ phonemes as represented by their hidden Markov models.

► **Step 0.** The number of independent models are $N_{it} = N$. N_{it} speech units are trained. Each of the speech units is being represented by approximately 200 randomly selected realisations taken from the combined training corpus.

► **Step 1.** For each model pair ϕ_i and ϕ_j of speech units, a value $c(\phi_i, \phi_j)$ is calculated on the basis of a recognition experiment in

which all stimuli segments - modelling all units $\phi_i, i \in \{1, \dots, N_{it}\}$ - of the combined test corpus is tested against all response segments - modelling units $\phi_j, j \in \{1, \dots, N_0\}$ - in a Viterbi search.

► **Step 2.** A similarity $S_{p,q}$ between any two speech units ϕ_p and ϕ_q is calculated on the basis of the following defined expression:

$$S(\phi_p, \phi_q) = \sum_{k=1}^{N_0} (c(\phi_k, \phi_p) - c(\phi_k, \phi_q))^2, \quad p \neq q$$

A total of $(N_0 * N_{it}) / 2$ calculations are performed.

► **Step 3.** The two closest speech units - say ϕ_p and ϕ_q are merged (meaning that the label representing sound unit ϕ_p is set equal to the label representing sound unit ϕ_q during re-annotating of the combined training corpus) into one common sound unit by averaging of their log-probability values:

$$c(\phi_k, \phi_p) = (c(\phi_k, \phi_p) + c(\phi_k, \phi_q)) / 2, \quad k \in \{1, 2, \dots, N_0\}$$

► **Step 4.** If a preselected number of equated phonemes have been identified go to Step 5. Else re-annotate the combined training corpus according to the merging of the two closest speech units, set $N = N - 1$ and go to Step 0. The merging continues until a selected number of combined, language-independent sound units are chosen by the iterative process.

► **Step 5.** The equated new speech units are defined as the group of *language-independent sound units*.

► **Step 6.** The remaining phonemes in each of the training corpora are defined as the group of *language-dependent phonemes* for each of the training languages.

It is emphasised that the data-driven, iterative process for selection of the language-independent sound units 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 sound units. This is utilised in experiment two.

6. PHONEME MODELLING

The phonemes are each modelled by a context independent hidden Markov model described by three states and with one skip. The statistics in each of the phoneme states is modelled by two mixtures.

Models of 'non-speech' segments are each described by an ergodic Markov model with four states and two mixtures in each state. The HTK-tool (version 1.5) is used for training.

7. LD-TRANSFORMATION

A linear discriminant (LD) analysis is introduced with the aim of decorrelating the input variables $\mathbf{x}(n)$ maximally, and the result is transformed into a set of new parameters $\mathbf{y}(n)$ which are subsequently used by the language classifier. Discriminant analysis involves deriving linear combinations of the input variables that will discriminate between a-priori defined classes in such a way that the misclassification error rates are minimised.

This is performed by determining a set of discriminant functions which best discriminate between the classes by maximising the ratio of the between-class variances to the within-class variances

subject to a number of constraints. Given that the following constraints are fulfilled, namely:

1. each class is a sample from a multivariate normal population,
2. all classes have identical within-class covariance matrices \mathbf{W} ,
3. the class means (centroids) can be represented by a Gaussian distribution with between-classes covariance matrix \mathbf{B} ,

then it is possible to define a transformation matrix \mathbf{E} and a feature reduction matrix \mathbf{f} which transforms the input parameters $\mathbf{x}(n)$ into an LD-transformed vector $\mathbf{y}(n)$ which best discriminates between the classes as given by the training data. A detailed derivation can be found in [9]. The result of the LD-transformation is that the language classifier is given the following input data:

$$\mathbf{y}(n) = \mathbf{f}^T \mathbf{E}^T \mathbf{x}(n).$$

8. LANGUAGE CLASSIFICATION

The transformed vector $\mathbf{y}(n)$ leads to the following expression

$$d_{ij} = \mathbf{y}^T (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j) - \frac{1}{2} (\bar{\mathbf{y}}_i + \bar{\mathbf{y}}_j)^T (\bar{\mathbf{y}}_i - \bar{\mathbf{y}}_j)$$

where the transformed mean class vectors \mathbf{y}_i and \mathbf{y}_j are computed on the basis of the training material. From this the final classification is estimated using Bayes' rule.

9. EXPERIMENTS

Two experiments are conducted. The first applies a total set of the $N = 113$ language-specific phoneme models, the second applies a set of varied-size groups of language-independent speech units and language-dependent phonemes. There are 40, 41 and 32 CDHMM phoneme models for the three languages American-English (US), German (GE) and Spanish (ES), respectively. In each of the experiments 10 'non-speech' models are used together with the speech models.

The first experiment analyses which of the three preprocessing techniques is optimal as regards the task of phoneme recognition, and which of the preprocessing techniques is optimal as regards the overall task of language-identification.

9.1 Experiment 1

The results from the phoneme recognition experiments are given in Table I.

The results show a relative low phoneme recognition performance which is explained by the fact that the simulations are performed on unconstrained spontaneous speech of telephone quality.

It is observed that FBDYN gives better performance than MFCC and RASTA.

Table II shows the results for language-identification.

The results show that the use of RASTA for preprocessing in combination with language models and LD-transformation give a substantial raise in language-identification score. The remaining

Table I. Phoneme recognition accuracy using MFCC, FBDYN and RASTA on *language-specific phoneme* models

	US	GE	ES	Average
MFCC	36.2 %	35.9 %	40.4 %	37.5 %
FBDYN	36.5 %	36.2 %	41.7 %	38.1 %
RASTA	36.1 %	34.9 %	42.2 %	37.7 %

Table II. Language-identification using MFCC, FBDYN and RASTA on *language-specific phoneme* models

	US	GE	ES	Average
MFCC	96 %	76 %	76 %	84 %
FBDYN	93 %	74 %	72 %	81 %
RASTA	96 %	90 %	89 %	92 %

experiments are conducted with only RASTA technique.

9.2 Experiment 2

The experiment is initialised on the basis of all $N_0 = 113$ phoneme models as trained in Experiment 1. First, 20 language-independent speech units have - arbitrarily - been chosen. The results of the experiment are shown in Table III.

Table III. Phoneme recognition accuracy using RASTA preprocessing and the combination of 20 *language-independent speech unit* and *language-dependent phoneme* models.

	US	GE	ES	Average
RASTA	36.4 %	34.6 %	41.9 %	37.0 %

Comparing with the results from the similar test in Experiment 1, it is observed that the phoneme recognition accuracy stays almost constant using 20 language-independent speech units together with language-dependent phonemes.

Table IV shows the results of testing language-identification scores of the baseline LID-system in which the number of language-independent speech units are varied.

It is seen that the overall performance remains almost constant despite the use of an increasing number of language-independent speech unit models in the recogniser, and it is observed that the language-identification score remains at a relative high level although the number of language-independent speech units used by the phoneme recognisers increases.

10. CONCLUSION

The results presented in this paper indicate that high LID-system

performance can be achieved for telephone quality speech by applying RASTA for preprocessing, by introducing clustering of phonemes across languages and by applying an LD-transformation

Table IV. Language-identification score using a varying number of *language-independent speech units* and *language-dependent phoneme* models

No. common speech units	US	GE	ES	Average
10	93 %	86 %	85 %	88 %
20	91 %	84 %	83 %	87 %
30	87 %	80 %	83 %	84 %
40	86 %	92 %	81 %	86 %
50	90 %	82 %	77 %	84 %

before final language classification takes place. Space does not allow results to be given on the influence of the LD-transformation. The technique presented in this paper enables testing of the LID-system in which the number of language-independent speech units is varied, and it is interesting to observe that this only affects the language-identification accuracy to a minor degree. It is presently being analysed whether the relatively small degrading in accuracy can be 'restored' by modelling the language by trigrams instead of bigrams as it is emphasised that an increase in number of language-independent speech units at the same time may allow for trigram language modelling.

11. REFERENCES

1. Zissman M A (1996). 'Comparison of Four Approaches to Automatic Language Identification of Telephone Speech'. IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 1, January, pp 31-44.
2. Yan Y, Barnard E, Cole R A (1996). 'Development of an approach to automatic language identification based on phone recognition'. Computer Speech and Language **10**, pp 37-54.
3. Davis S D, Mermelstein P (1990). 'Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuous Spoken Sentences'. Readings in Speech Recognition, Edited by A Waibel, K-F Lee
4. Hermansky H, Morgan N, Bayya A, Kohn P (1991). 'Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech'. Int. Conf. EUROSPEECH'91, pp 1367-1370.
5. Beppu T, Aikawa K (1995). 'Spontaneous Speech Recognition Using Dynamic Cepstra Incorporating Forward and Backward Masking Effect', Int. Conf. EUROSPEECH95, pp 511-514.
6. Muthusamy Y K, Cole R A, Oshika B T (1992). 'The OGI multi-language telephone speech corpus', Int. Conf. ICSLP92, pp 895-898.
7. Dalsgaard P, Andersen O (1994). 'Application of Inter-Language Phoneme Similarities for Language Identification', Int. Conf. ICSLP94, pp 1903-1906.
8. Andersen O, Dalsgaard P, Barry W (1993). 'Data-Driven Identification of Poly- and Mono-phonemes for four European Languages', Int. Conf. on EUROSPEECH93, pp 759-762.
9. Morrison D F (1990). 'Multivariate Statistical Methods', Third edition, McGraw Hill, ISBN 0-07-043187-6.