LANGUAGE IDENTIFICATION WITH LANGUAGE-INDEPENDENT ACOUSTIC MODELS*

C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{corredor,gauvain,madda,lamel}@limsi.fr

ABSTRACT

In this paper we explore the use of language-independent acoustic models for language identification (LID). The phone sequence output by a single language-independent phone recognizer is rescored with language-dependent phonotactic models approximated by phone bigrams. The language-independent phoneme inventory was obtained by Agglomerative Hierarchical Clustering, using a measure of similarity between phones. This system is compared with a parallel language-dependent phone architecture, which uses optimally the acoustic log likelihood and the phonotactic score for language identification. Experiments were carried out on the 4-language telephone speech corpus IDEAL, containing calls in British English, Spanish, French and German. Results show that the language-independent approach performs as well as the language-dependent one: 9% versus 10% of error rate on 10 second chunks, for the 4-language task.

1. INTRODUCTION

This paper presents some of our recent research on automatic language identification (LID) for telephone speech. Previous work on LID has shown that the most accurate approaches are based on language-dependent phone recognizers used in parallel[8, 9]. In approaches based on the joint acoustic-phonotactic log likelihood, the identified language is associated with the model set having the highest score[6, 7, 8, 9], and the actual phone sequence produced by the system is ignored. In contrast, in a purely phonotactic approach, the acoustic scores are ignored and the phone sequence output from each model set serves as input to the phonotactic n-gram models[8, 9].
Although parallel architectures appear to be more accurate, they exhibit two important problems: they need phonetically labeled training data for each target language,[1] and they can require relatively important decoding times. An approach based on a single language-independent phone recognizer reduces these problems, but the choice of the common phoneme set is important. This is the main problem we have investigated in this work. We have also extended our parallel language-dependent phone recognition architecture to optimally use both the acoustic and phonotactic information for LID[3]. The approaches were evaluated on IDEAL, a multi-language corpus containing telephone speech in British English, Spanish, French and German.

2. THE IDEAL CORPUS

The IDEAL corpus is a multi-language telephone speech corpus designed to support research on LID. This corpus offers several advantages as compared to other multi-language corpora: it contains a large amount of speech (about 19 hours per language), the different languages were collected under the same conditions, and native speakers were recruited in their home countries. At present, data have been recorded for British English, Spanish, French and German. The IDEAL corpus contains about 300 "*matched calls*" for each language (i.e., native U.K., Spanish, French and German speakers calling from their home country), and up to 70 "*crossed calls*" for each language (i.e., native French speakers calling from the U.K., Germany and Spain, and native U.K., Spanish, and German speakers calling in France). All speakers called the LIMSI data collection system free of charge, assuring the same recording conditions for the entire corpus.
The calling script, slightly modified to fit each language and country, was designed to cover a variety of data types:

- 12 questions to elicit responses: 7 general questions concerning the call and caller (code, sex, age, mother language, city name, postal code, and the first digits of their phone number), and 5 prompts asking for times, dates, days of the week and months of the year ("what time is it now?", "what is today's date?", "what is the birthday of someone you know?", etc.).

- 18 items containing predefined texts to read (2 newspaper sentences, 2 travel-related sentences, 2 phonetically rich sentences, one information request, dates, times, spoken and spelled words,

* THIS WORK WAS PARTIALLY FINANCED BY A CNET CTI PROJECT.

[1] In phonotactic approaches, phonetically labeled training data for each target language is not a requirement, but the system performs better when labeled data is available for each of the languages[9].

phone numbers, credit card numbers, spoken and spelled names, digit strings, addresses, and money amounts).

- 6 questions aimed at collecting spontaneous speech ("What is your dream holiday?", "What type of restaurant do you prefer?", etc.).

250 of the "*matched calls*" (about 9000 sentences, containing up to 13 hours of speech for each language) have been used for training material. Two test corpora were defined: the IDEAL "*matched test*" corpus, including utterances from about 50 "*matched calls*"; and the IDEAL "*crossed test*" corpus, with sentences from up to 50 "*crossed calls*". The evaluations reported here were carried out on the 6 free spontaneous speech utterances, as this data is assumed to be the most representative of the type of data expected in a real application.

3. THE LANGUAGE-DEPENDENT APPROACH

Let $L = \{L_1, L_2, \ldots, L_N\}$ the set of languages to be identified. The approach based on language-dependent phone recognition uses a bank of N parallel phone recognizers followed by phonotactic bigram models (N models per phone recognizer, as shown in Figure 1).

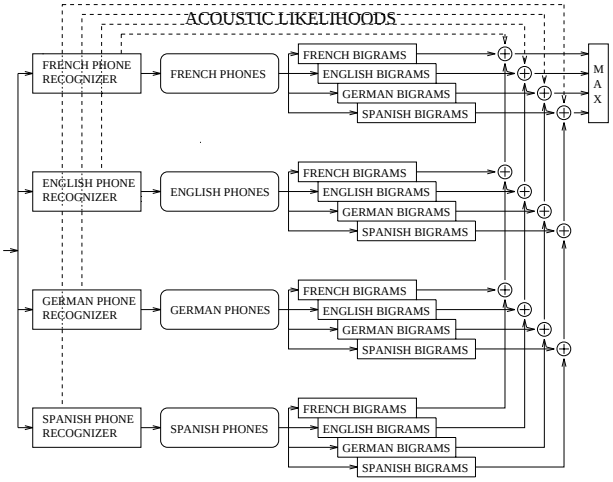

Figure 1: Block diagram of the parallel language-dependent phone recognition approach to LID.

The incoming test utterance is decoded by all language-dependent phone recognizers. Phonotactic constraints, provided by an *embedded* phonotactic bigram model, are applied during the Viterbi process of each recognizer. The L_i embedded phonotactic model differs from the N phonotactic models (*secondary* bigram models) used to compute the phonotactic score of the output of L_i recognizer. The embedded bigram model is trained on the phone transcriptions of the L_i training corpus, where the phone labels are obtained by forced alignment. The N L_i secondary bigram models are also estimated on the same training data, but use the output of the corresponding recognizer.

A language-dependent score ℓ_i is obtained as the sum of two terms[3]

$$\ell_i = \log f(\vec{a}|L_i) + \sum_{j=1}^{N} \frac{1}{T_j} \log \Pr(\Phi_j|L_i) \qquad (1)$$

where $\vec{a}$ is the acoustic representation of the test utterance, L_i is a language, $\Phi_j = \{\varphi_1, \varphi_2, \ldots, \varphi_{T_j}\}$ is the phone sequence output from the L_j phone recognizer, and T_j is the number of recognized symbols. The first term in (1) is the output acoustic log likelihood (normalized by utterance length) of the L_i phone recognizer. The average log likelihood over all utterances decoded by this phone recognizer is subtracted from this score to remove any bias between phone recognizers[9]. The second term in (1) is the phonotactic score, computed from phone string output by the N phone recognizers. The phonotactic score is the sum of the corresponding log probabilities (normalized by number of phones in the utterance) for each of the N phone bigram models. The language with the highest score ℓ_i is hypothesized.

The language-dependent phone recognizers

Acoustic models were built for British English, Spanish, French and German phones using the IDEAL training corpus. Each phone of each language-dependent recognizer is modeled by a 3 state context-independent CDHMM. The quality of these models was evaluated on about 200 utterances per language from the "*matched test*" subcorpus. The results, in terms of phone recognition error, are shown in Table 1.

Language	*# of phones*	*Error*
English	44 + *sil*	67.3%
Spanish	24 + *sil*	48.7%
French	34 + *sil*	55.7%
German	47 + *sil*	56.9%

Table 1: Phone recognition results (% error) on 200 utterences per language from the IDEAL "*matched test*" corpus, using the given number of phones and the corresponding phone bigrams.

4. THE LANGUAGE-INDEPENDENT APPROACH

The approach based on language-independent phone recognition uses one single phone recognizer to label the speech input. The recognized phone string is rescored using N phonotactic bigram models[6] (see Figure 2), and the language providing the highest log probability is hypothesized. For each language, the bigram model was estimated on the labels output by the language-independent phone recognizer on the training data for that language.

Other decoding strategies are possible using a single set of language-independent acoustic models. This is the most practical implementation with respect to decoding and training, in particular to simplify extension to more languages.

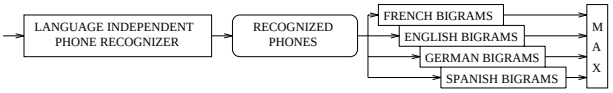

Figure 2: Block diagram of the language-independent phone recognition approach to LID.

The language-independent phoneme inventory

To obtain the language-independent phoneme set, all the phones of each of the 4 languages were clustered using an Agglomerative Hierarchical Clustering algorithm[1, 2]. One of the problems associated with clustering techniques is the choice of a reasonable similarity (or dissimilarity) measure between samples (between phones in our application). In previous work, the dissimilarity measure between phones φ_i and φ_j was defined as the log likelihood difference of the phone observation sequence $\vec{\varphi}_i$ conditioned on the λ_i and λ_j HMMs[4, 5]. In this study, we use a similarity measure $S(\varphi_i, \varphi_j)$ which approximates the *a posteriori* probability $\Pr(\lambda_i|\vec{\varphi}_j)$:

$$S(\varphi_i, \varphi_j) = f(\vec{\varphi}_j|\lambda_i)^\gamma / \sum_{i=1}^{n} f(\vec{\varphi}_j|\lambda_i)^\gamma \qquad (2)$$

where $\vec{\varphi}_i, \vec{\varphi}_j$ are the labeled training data corresponding to phone φ_i and φ_j; λ_i, λ_j are the associated CDHMMs, and n is the number of units to be clustered. The normalization coefficient γ (needed to compensate for independency approximations in the models) was empirically determined to be 0.5. We use a symmetrized version of (2)

$$S_s(\varphi_i, \varphi_j) = \frac{S(\varphi_i, \varphi_j) + S(\varphi_j, \varphi_i)}{2} \qquad (3)$$

Another important issue is the definition of the similarity measure between two clusters. This inter-cluster measure corresponds to the two-phone similarity when there is only one phone per cluster. In the case of more than one phone per cluster, the similarity measure between cluster C_i and C_j is defined as

$$S(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\varphi \in C_i} \sum_{\varphi' \in C_j} S_s(\varphi, \varphi') \qquad (4)$$

where n_i and n_j are the number of phones in C_i and C_j respectively.

The hierarchical clustering procedure was applied to the labeled training data for all 4 languages. The initial 148 language-dependent phonemes were grouped into 83 clusters, of which 48 are singletons, that is each corresponds to a language dependent phoneme. The remaining 35 clusters contain multiple language-dependent phonemes as shown in Table 2.

The 48 singletons are distributed across languages as follows: 15 for British English, 6 for Spanish, 9 for French, and 18 for German. These account for 25% to 40% of the phonemes in each language. Some examples of these language-dependent units are: /ð/,/θ/,/ʌ/ for British English; /R/,/x/,/ʧ/ for Spanish, /œ̃/, /ɥ/,/œ/ for French; and /ç/,/x/,/Y/,/ø/ for German.

Consonants				*Consonants*			
En.	*Sp.*	*Fr.*	*Ge.*	*En.*	*Sp.*	*Fr.*	*Ge.*
p	p	p	p	–	–	r	r
t	t	t	t	z,v	–	–	–
k	k	k	k	h	–	–	h
g	–	g	g	–	ʎ̟,ʝ̟	–	–
–	–	b,d	–	*Vowels*			
b	–	–	b	æ	a	a	a,ɑ
d	–	–	d	i	i	i	i
–	b,d	–	–	–	–	e	e:,e,I
n	n	n	n,ə:n	ɔ:	o	ɔ	ɔ
m	m	m	m	–	e	ɛ	ɛ
–	ɲ	ɲ	–	ɔ	–	o	o
ŋ	–	–	ŋ	eə, ɛ,ɜ˞	–	–	–
ʧ,dz	–	–	–	–	u	u	–
s	s	s	s	–	–	–	ʊ,u
f	f	f	f	I,e	–	–	–
ʃ	–	ʃ	ʃ	ɔi	–	–	ɔ
ʒ	–	ʒ	–	–	–	$\tilde{a}$,$\tilde{e}$	–
–	l	l	l	–	–	y	y

Table 2: Multi-phoneme clusters obtained by agglomerative hierarchical clustering. Each row corresponds to a cluster, where the language-dependent phonemes are represented using IPA symbols. The columns correspond to the source languages, *En.*: British English; *Sp.*: Spanish; *Fr.*: French; *Ge.*: German.

Of the non-singleton clusters, 22 regroup 62 language-dependent consonants and the remaining 13 regroup 38 language-dependent vowels. Only 5 are purely *intra-language* clusters (French /b/,/d/; Spanish /b/,/d/; English /ʧ/,/ʤ/; English /z/,/v/; and Spanish /ʎ̟/,/ʝ̟/) which correspond mainly to voiced plosive and fricative sounds. Note that the pair of sounds /ʎ̟/,/ʝ̟/ corresponds to a well-known Spanish problem[2]. Most speakers pronounce the sound /ʎ̟/ as /ʝ̟/, which explains why these phone units were clustered into the same group. Seven clusters group together the same phone label across the 4 languages: /p/,/t/,/k/,/s/,/f/ (unvoiced plosives and fricatives), and /n/,/m/ (nasals). The automatically derived clusters for the most part correspond to linguistically similar or identical units.

For the non-singleton clusters, phoneme cluster models are trained using all the training data labeled with the corresponding language-dependent labels. A common silence model was added to the 83 language-independent model set.

5. EXPERIMENTS

The results of experiments carried out on 5s and 10s chunks from the IDEAL "*matched test*" corpus are given in Table 3. The approach based on language-dependent phone recognition (using both the acoustic and the phonotactic scores, LD_AP) was evaluated and compared to results obtained without integration of the acoustic log likelihood (LD_P, called par-

[2]called "*yeismo*" in Spanish.

allel PRLM by Zissman[9]). The results show that by adding acoustic information to the phonotactic score, a 17% of relative error reduction on 10s chunks can be obtained (10% of absolute error rate for LD_AP versus 12% of absolute error rate for LD_P). Comparing LD_AP to results obtained using only the acoustic log likelihood (LD_A, called PPR by Zissman[9] and Kwan[6]), a 44% of relative error reduction on 10s chunks was observed. The approach based on language-independent phone recognition (using the hierarchically clustered phone set, LI_HC) is more accurate than either the LD_A or the LD_P approaches. Comparing results to LD_AP, only a slight degradation on 5s chunks was noted; whereas a 10% of relative error reduction on 10s chunks was obtained.

Approach	*5s chunks*	*10s chunks*
LD_AP	15%	10%
LD_P	22%	12%
LD_A	20%	18%
LI_HC	18%	9%

Table 3: LID error rates on 5s, 10s "*matched test*" chunks for 4-language task. LD_AP: language-dependent, acoustic and phonotactic model; LD_P: language-dependent, phonotactic model; LD_A: language-dependent, acoustic model; LI_HC: language-independent, hierarchicaly clustered phoneme set.

The above approaches were also evaluated on the "*crossed test*" subcorpus, as shown in Table 4. A notable degradation was observed for all approaches, that can be attributed to the acoustic mismatch between training and testing conditions, and/or due to different characteristics of speech from speakers recruited outside of their native-language country. For example, speakers abroad are more likely to use foreign words in their responses than speakers calling from their native country. The approaches using phonotactic scores (LD_P and LD_HC) appear to be more robust than the approaches based on acoustic log likelihoods (LD_A and LD_AP). When the acoustic score is used, the error rate is doubled for the "*crossed test*". The language-independent approach (LI_HC) has only a small performance degradation compared to the best language-dependent approach (LD_P): 29% versus 28% on 5s chunks, and 21% versus 17% on 10s chunks (absolute error rates).
Comparing the results on the two test subcorpora, the high degradation observed for the LD_A approach (18% versus 41% absolute error rate on 10s chunks) suggests that the language-dependent phone recognizers are simultaneously modeling the acoustic-phonetic information and the characteristics of telephone network, despite the use of cepstral mean removal. Thus relatively optimistic results are obtained in the "*matched test*" subcorpus, where both the language and channel characteristics are the same for the training and test data, and relatively pessimistic results are obtained for the "*crossed test*" data. For each approach, more realistic error rates are expected to fall at some point in between these two results.

Approach	*5s chunks*	*10s chunks*
LD_AP	33%	27%
LD_P	28%	17%
LD_A	43%	41%
LI_HC	29%	21%

Table 4: LID error rates on 5s, 10s "*crossed test*" chunks for 4 language task. LD_AP: language-dependent, acoustic and phonotactic model; LD_P: language-dependent, phonotactic model; LD_A: language-dependent, acoustic model; LI_HC: language-independent, hierarchicaly clustered phoneme set.

6. CONCLUSION

In this paper we have described a method for automatic language identification using a single language-independent phone recognizer. A hierarchicaly clustering algorithm was used to group sets of language-dependent phonemes. The resulting clusters often correspond to linguistically similar or identical phonemes of the different languages. Experiments using the IDEAL corpus indicate that the language-independent approach performs as well as the best of the other 3 approaches using language-dependent acoustic models. Moreover, the language-independent approach offers several advantages for extension to new languages. First, once the language-independent acoustic model set has been built, there should be no need for labeled training data, at least for close languages. Second, the only additional computation needed to add a new language is to label the training data in order to estimate the phonotactic model. Third, the decoding time essentially remains the same when a new language is added.

REFERENCES

[1] K.M. Berkling, E. Barnard, "Language Identification of six Languages Based on a Common Set of Broad Phonemes", *ICSLP-94*.

[2] R.O. Duda, P.E. Hart, "Pattern Classification and Scene Analysis", *Wiley-Interscience*, 1973.

[3] T.J. Hazen, V.W. Zue, "Recent Improvements in an Approach to Segment-based Automatic Language Identification", *ICSLP-94*.

[4] B.H. Juang, L.R. Rabiner, "A Probabilistic Measure for Hidden Markov Models", *AT&T Technical Journal*, **64**(2), Feb., 1985.

[5] J. Köhler, "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds", *ICSLP-96*.

[6] H.K. Kwan, K. Hirose, "Recognized Phoneme-based n-gram Modeling in Automatic Language Identification", *Eurospeech'95*.

[7] L.F. Lamel, J.L. Gauvain, "Language Identification Using Phone-based Acoustic Likelihoods", *ICASSP-94*.

[8] Y.K. Muthusamy, E. Barnard, R.A. Cole, "Reviewing Automatic Language Identification", *IEEE Signal Processing Magazine*, Oct., 1994.

[9] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", *IEEE Trans. on SAP*, **4**(1), Jan., 1996.