LANGUAGE IDENTIFICATION INCORPORATING LEXICAL INFORMATION

D. Matrouf, M. Adda-Decker, L.F. Lamel, J.L. Gauvain

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{matrouf,madda,lamel,gauvain}@limsi.fr

ABSTRACT

In this paper we explore the use of lexical information for language identification (LID). Our reference LID system uses language-dependent acoustic phone models and phone-based bigram language models. For each language, lexical information is introduced by augmenting the phone vocabulary with the N most frequent words in the training data. Combined phone and word bigram models are used to provide linguistic constraints during acoustic decoding. Experiments were carried out on a 4-language telephone speech corpus. Using lexical information achieves a relative error reduction of about 20% on spontaneous and read speech compared to the reference phone-based system. Identification rates of 92%, 96% and 99% are achieved for spontaneous, read and task-specific speech segments respectively, with prior speech detection.

1. INTRODUCTION

Many state-of-the-art language identification (LID) systems exploit phone-based acoustic and (or) phonotactic scores [7]. Training generally consists of designing one phone-based recognizer per language (i.e., there is no explicit use of lexical information). During test, these recognizers are run in parallel, and the one with the highest likelihood is selected, with the language associated with the model set identified [2].

Theoretically, if a large vocabulary continuous speech recognition system (LVCSR) was substituted for the phone-based system in each language, better language identification results could be achieved. This is because LVCSR systems use higher level knowledge: words and sequences of words rather than phonemes and phoneme sequences. In practice this approach has not been widely explored [4], since in addition to being computationally expensive, it is difficult to use if only small amounts of language-specific data are available.

The words in a language are not evenly distributed – the most frequent words account for a large proportion of all word occurrences. For large newspaper corpora in English (*WSJ*) and French (*Le Monde*), the most frequent 100 words account for about 40% of all word occurrences. For task specific vocabularies (such as travel information tasks) the lexical coverage for the 100 most frequent words is about 70%. This property may be taken advantage of in building a system for language identification.

	Lexical Coverage (%) of N words					
Language (M)	10	50	100	250	500	1K
English (2341)	27	49	60	72	82	91
French (2400)	26	54	64	76	84	91
German (3255)	22	44	57	68	77	86
Spanish (5008)	28	52	61	72	79	86

Table 1: Lexical coverage rates (%) of spontaneous training data in the IDEAL corpus for the N most frequent words. For each language the number of distinct words M in the spontaneous training data is also given.

In this contribution we address the following interrelated questions:

- To what extent do lexical constraints improve LID?
- Is LID easier for task-specific domains than for more general topics?
- Is LID more difficult with spontaneous speech than with read or elicited speech?

In the next section we describe our new strategy combining a phoneme-based models with lexical information from the most frequent words. Section 3 describes speech corpus and presents experimental results for different training and test configurations. The experimental setup was designed to give at least partial answers to all of the questions stated above.

2. USE OF LEXICAL INFORMATION

The motivation for incorporating lexical information the acoustic approach stems from the observation that relatively high lexical coverages can be achieved using a relatively limited number of words. Table 1 shows the lexical coverage rates obtained for different values of the N most frequent words in the spontaneous speech portion of the 4-language IDEAL corpus [3]. The 10 most frequent words account for about 25% of all word occurrences in the training data, and about 70% of the training data are covered using $N = 250$. These figures hold approximately for the

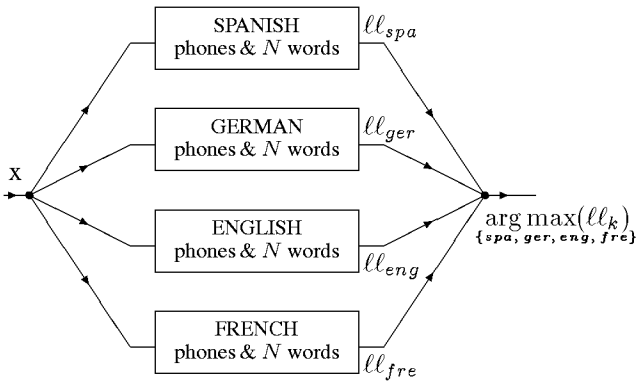

Figure 1: Block diagram of the parallel language-dependent **phone & N most frequent word** recognition approach to LID.

four languages studied, despite the differences in the total number of distinct words in the transcriptions (M).[1].

The approach described here is an extension of the parallel phone recognition approach used in [2], [7], where instead of modeling linguistic information only by phonotactic constraints, for each language the N most frequent words are also taken into account.

Let $L = \{L_1, L_2, \ldots, L_K\}$ the set of languages to be identified. The approach based on language-dependent phone recognition uses a bank of K phone recognizers, with a specific phone set for each language. Acoustic models are trained for each language k and language model constraints are provided by phone bigrams.[2]

In the proposed approach the acoustic models remain unchanged, but each system vocabulary contains its language-specific phones and the N most frequent words observed in training data for the language. The orthographic transcriptions of the training are transformed to replace all words not in the N most frequent words by their phone transcriptions (obtained by Viterbi alignment). The resulting transcripts, consisting of sequences of phones and words are used to estimate hybrid language models using standard estimation techniques.

The system architecture is shown in Figure 1, where the incoming test utterance x is decoded by the K language-dependent **phone & N most frequent word** recognizers. Some example system outputs are shown in Figure 2. In the first example the system outputs mostly words. In the second sentence, the unknown word "girl" (followed by "and") is replaced with "garden" and the unknown word "boys" is recognized phonemically. The third example is recognized as a mix of words and phones.

Each of the K recognizers produces a log-likelihood $\ell\ell_k$ which is used to take the LID decision. In our present system this is simply the maximum likelihood criterion.

[1] The significantly higher number for Spanish ($M = 5008$) is due to the larger amount of spontaneous speech collected: for the same number of responses, twice as much speech data was collected for the Spanish language as compared to German, English or French.

[2] In the parallel approach it is common to use sets of phonotactic bigram models to rescore the parallel outputs [7], offering the advantage of being able to identify languages for which only untranscribed training data is available. This work does not use any subsequent phonotactic bigram models.

Language	*#Calls*	*#Male*	*#Female*	*#Hours*
English	258	109	149	14.8
French	259	129	130	13.1
German	257	109	148	15.8
Spanish	253	114	139	17.9

Table 2: Summary of data under matched language/country conditions.

3. LID EXPERIMENTS

Experiments have been carried out to assess the contribution of lexical information on 3s and 5s segments of the 4-language telephone speech corpus IDEAL [3]. Automatic language identification research using this corpus has been reported in [1].

3.1. The IDEAL telephone speech corpus

IDEAL is a large, four-language corpus (French, British English, German and Castillan Spanish) of telephone speech for research in automatic language identification [3]. The corpus is similar in style to the OGI multi-language corpus [5], containing read and spontaneous speech for each caller. The corpus contains data from over 250 native speakers of each language calling from their home country (matched language/country conditions), and an additional 50 calls per language from another country (crossed conditions). Table 2 summarizes the matched data for the different languages.

The callers, balanced for sex, age and dialect, were recruited by a marketing survey company who distributed calling designed to collect three types of data:

- **Call information:** general questions concerning the call and caller, these data were not used in these experiments.
- **Read & elicited speech:** items containing pre-defined texts to read and fixed prompts ("what time is it now?");
- **Spontaneous speech:** a set of questions aimed at obtaining spontaneous speech ("speak about your home, your dream vacations, your favorite music" etc.)

The **read and elicited** speech items in the caller scripts were generated automatically from source files containing several thousand different texts for each item. These include texts extracted from newspapers, simple telephone introductory phrases or information requests, travel information queries, dates, times, credit card numbers, telephone numbers, spoken and spelled common words and proper names, digit strings, money amounts, and complete names and addresses. The high proportion of items including **numbers** and dates motivated the LID test on these data

T: having to **wait** uh for long **periods** for the **bus** to **come** as it's **late** on it's **schedule** and so on
Hyp: I having to way to prefer shopping for the carpet and chips and they car etc and a n t

T: 3 children 1 **girl** and 2 **boys**
Hyp: 3 children 1 garden to b c I z

T: the last time I went to a museum was the **sea** life centre and we saw lots of **various** fish in **their natural surroundings**
Hyp: f @ t W Y n to museum was to see my friends k l for lots of b R l u fish and then @ C r look for and I G k s

Figure 2: Some example output showing the partial hypotheses. The words in transcript *T* that are shown in bold are not in the recognition lexicon.

(see below). The **spontaneous** portion of the corpus contains responses to a series of questions selected randomly at record time from a set of about 200 questions. The questions were not written on the paper script, in order to prevent callers from preparing their answers. The spontaneous data accounts for about 15% of the corpus, not including silences.

3.2. Experimental conditions

Specific test sets were selected so as to be able to compare LID performance on spontaneous speech to read/elicited speech. Two different sets of data were used for read and elicited speech. The first set included all read and elicited items (i.e. newspaper texts, travel information queries, dates, numbers, addresses ...). The second is a subset of this data including only items related to numbers and dates. The lexical information was included by adding the N most frequent words in the respective subcorpa of the training data: spontaneous speech transcripts, read & elicited speech transcripts, and number & date transcripts. Bigram language models were trained for each test condition: *spontaneous, read, numbers*.

The same set of acoustic models were used for all experiments. These models were trained on all of the training data (spontaneous and read speech) from 200 calls per language. 50 calls per language were reserved for test.

For the test condition ($\mathcal{T}_{condition}$), all utterances of the *condition* with a minimal duration (5s or 3s) were used. Only the first part of the acoustic signal of each utterance was used for the LID test. To investigate the extent to which the LID results are influenced by non-speech acoustic segments, an additional series of tests were carried out using prior speech detection, where $\mathcal{T}'_{condition} \subset \mathcal{T}_{condition}$. Speech detection was obtained by aligning the data with the transcripts, simulating optimal speech/non-speech detection. After removing initial and final silence portions, the $\mathcal{T}'_{condition}$ test set contains the speech segments containing at least 5s of speech. In future work we will measure the effect of using an automatic algorithm for speech/non-speech detection (i.e. without using the transcriptions).

3.3. Spontaneous speech

The N most frequent words and the hybrid language models are obtained exclusively from the spontaneous speech portion of the training corpus. The test data $\mathcal{T}_{spont}$ (871 5s segments) and $\mathcal{T}'_{spont}$ (588 5s segments) also contain only spontaneous speech. The lexical coverage of the spontaneous training data were shown to be somewhat comparable for different languages (see Table 1). In Table 3 the lexical coverage rates, averaged across languages, are given for both training and test data. The difference in coverage between training and test is small for all values of N, but increases with N.

	Lexical coverage		*%LID error*	
	Train	*Test*	$\mathcal{T}_{spont}$	$\mathcal{T}'_{spont}$
N	*#of 5s segments*		871	588
0	-	-	17.0	11.6
100	60.3	59.4	13.8	9.2
250	72.0	70.4	13.4	8.3
500	80.6	78.3	12.4	8.0
N	*# of 3s segments*		1242	840
0	-	-	21.0	16.2
100	60.3	59.4	17.9	13.0
250	72.0	70.4	16.9	11.8
500	80.6	78.3	15.9	11.3

Table 3: LID approach combining phonemes and N most frequent words for LM. Language identification error rates on 5s segments (top) and 3s segments (bottom) of **spontaneous** speech for the 4-language task as a function of N. Results are given without speech detection $\mathcal{T}_{spont}$ and with prior speech detection $\mathcal{T}'_{spont}$.

Table 3 shows the language identification error rates for different values of N on $\mathcal{T}_{spont}$ and $\mathcal{T}'_{spont}$. The LID error rates for $N = 0$ correspond to the phone-only approach. Incorporating lexical knowledge by including only a relatively small number ($N = 100$) of frequent words is seen to improve the relative performance by 15 to 20%. The performance improvement is larger on the set of segments with speech detection $\mathcal{T}'$.

Including more words ($N = 250, 500$) results in further performance gains. A relative error reduction of over 10% is observed by increasing N from 100 to 500. Comparing $\mathcal{T}_{spont}$ and $\mathcal{T}'_{spont}$ error rates for the 5s segments, speech detection results in a relative gain of more than 30% for all values of N. For the 3s segments, the difference in performance is over 20%. The 3s results with speech detection are seen to be better than the 5s results without. These differences highlight the importance of properly handling non-speech segments in optimizing LID systems.

3.4. Read and elicited speech

We investigated the performance on the read and elicited speech parts of the IDEAL corpus in order to measure the

	Lexical coverage		*%LID error*	
N	*Train*	*Test*	$\mathcal{T}_{read}$	$\mathcal{T}'_{read}$
0	-	-	7.9	5.4
100	72.3	72.0	5.7	4.8
350	85.7	85.1	5.0	4.2
500	88.3	87.5	5.0	4.0

Table 4: Language identification error rates on 5s segments of **read and elicited** speech for the 4-language task as a function of N. Results are without ($\mathcal{T}_{read}$) and with prior speech detection ($\mathcal{T}'_{read}$). $\mathcal{T}_{read}$: 1409 5s segments, $\mathcal{T}'_{read}$: 644 5s segments.

	Lexical coverage		*%LID error*	
N	*Train*	*Test*	$\mathcal{T}_{numbers}$	$\mathcal{T}'_{numbers}$
0	-	-	3.6	1.9
100	97.1	96.9	3.0	0.6
250	99.8	99.5	2.0	0.3

Table 5: Language identification error rates on 5s segments of read and elicited speech concerning the **numbers** domain for the 4-language task as a function of N. Results are given for 5s segments on $\mathcal{T}_{numbers}$ (no prior speech detection, #of 5s segments: 642) and on $\mathcal{T}'_{numbers}$ (prior speech detection, #of 5s segments: 321).

impact of a more carefully produced speech on LID rates. Read speech is known to be. on the average, more clearly articulated than spontaneous speech, with a lower rate of speaker produced noises such as breath and hesitations. Results are given in Table 4, where the $\mathcal{T}_{read}$ test set is comprised of 1409 5s speech segments and the $\mathcal{T}'_{read}$ test set contains about 644 5s segments of speech.

The use of lexical knowledge reduces the LID error by 28% ($N = 100$) for the ($\mathcal{T}_{read}$) test set without no prior speech detection. Using more words ($N = 350$) reduces the LID error by an additional 10%. However, despite the slightly higher lexical coverage with 500 words, the LID performance is not improved. Similar improvements were observed with 3s segments of speech.

A similar observation can be made for the ($\mathcal{T}'_{read}$) test set of 5s speech segments after speech detection. The LID error rate achieved by the acoustic phone-based approach ($N = 0$) is 5.4% and can be reduced to 4.3% by incorporating lexical knowledge about the 500 most frequent words. This corresponds to a 20% relative error reduction.

3.5. Task-oriented read and elicited speech

Here we consider a subcorpus of the read and elicited speech part of the IDEAL corpus to measure the impact of using a limited, task-specific vocabulary. The subcorpus consists of items containing mostly numbers: dates, times, credit card and telephone numbers, digit strings and money amounts. As can be seen in Table 5 very high lexical coverages can obtained on this type of data: 100 words cover 97% of all word occurrences, and 250 words covers over 99%.

$\mathcal{T}_{numbers}$ is a test set composed of 642 5s segments, and the $\mathcal{T}'_{numbers}$ test subset contains 321 5s segments of speech.

The LID errors rates are significantly lower than those obtained for more general tasks (compare this table with Tables 3 and 4. With prior speech detection, the LID rate is close to 100% on the 5s segments ($\mathcal{T}'_{numbers}$). Significant gains are still observed by increasing N, with the LID error for $N = 250$ being half that of $N = 100$.

These results clearly show the impact of linguistic content on LID rates. Even for the phone-based approach, the task-specific phone bigram, used during the acoustic Viterbi search, can capture some of this information.

4. CONCLUSIONS & PERSPECTIVES

In this paper we have experimented with an alternative approach for automatic language identification which makes combined use of phonemic and lexical information. This approach is an extension of the parallel language-dependent phone-based acoustic decoders, which are augmented by the N most frequent words of the given language. Incorporating lexical information yields a relative error reduction of about 15-30% depending upon the condition. For a given condition, LID rates were shown to increase with increasing lexical coverage. Since lexical coverages are typically higher in specific domains, better LID can be expected. The LID error for spontaneous speech (13.4%) is more than twice as high as for read speech (5.7%) given comparable lexical coverages of about 70%. A substantial reduction in error rate was obtained by removing initial and final non-speech portions of the signal. These non-speech events represent a noise source for the LID process, which is not sufficiently accounted for language-independent acoustic silence and noise models.

5. REFERENCES

1. C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel, "Language Identification with Language-Independent Acoustic Models," *Eurospeech'97*, **1**, pp. 55-58, Rhodes, Sept. 1997.
2. L. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *Eurospeech'93*, Berlin, **1**, pp. 23-28, Sept. 1993.
3. L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J.J. Gangolf, J.L. Gauvain, "A Multilingual Corpus for Language Identification," *1st International Conference on Language Resources and Evaluation*, **1**, pp. 1115-1122, Granada, May 1998.
4. S. Lowe, A. Demedts, L. Gillick, M. Mandel, B. Peskin, "Language Identification via Large Vocabulary Speaker Independent Continuous Speech Recognition," *ARPA Human Language Technology Workshop*, pp. 437-441, Plainsboro, March 1994.
5. Y.K. Muthusamy, R.A. Cole, B.T. Oshika (1992),"The OGI Multi-Language Telephone Speech Corpus," *ICSLP-92*, **2**, pp. 895-898 Banff, Oct. 1992.
6. T. Schultz, A. Waibel, "Fast Bootstrapping of LVCSR Systems with Multilingual Phone Sets," *Eurospeech'97*, **1**, pp. 371-374, Rhodes, Sept. 1997.
7. M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. on SAP*, **4**(1), Jan. 1996.
8. M.A. Zissman, "Predicting, Diagnosing and Improving Automatic Language Identification Performance," *Eurospeech'97*, **1**, pp. 51-54, Rhodes, Sept. 1997.