Language Identification Using Phone-based Acoustic Likelihoods

L.F. Lamel and J.L. Gauvain
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{lamel,gauvain}@limsi.fr

ABSTRACT

In this paper we apply the technique of phone-based acoustic likelihoods to the problem of language identification. The basic idea is to process the unknown speech signal by language-specific phone model sets in parallel, and to hypothesize the language associated with the model set having the highest likelihood. Using laboratory quality speech the language can be identified as French or English with better than 99% accuracy with only as little as 2s of speech. On spontaneous telephone speech from the OGI corpus, the language can be identified as French or English with 82% accuracy with 10s of speech. The 10 language identification rate using the OGI corpus is 59.7% with 10s of signal.

INTRODUCTION

Automatic language identification has a wide range of applications in providing voice access to a variety of computer and telephone-based services. For example, at information centers in public places, such as train stations and airports, the language may change from one user to the next. Under these conditions, it would be advantageous to be able to recognize the spoken query without prior knowledge of the language being spoken. Automatic language identification avoids having to ask the user to select the language before beginning to interrogate the system. Language identification has many other potential uses including: emergency situations (people in stressed conditions will tend to speak in their native tongue, even if they have some knowledge of the local language); travel services; communications related applications (translation services, information services, etc.); as well as the well-known national security applications.

While automatic language identification has been a research topic for over 20 years, there are relatively few studies published in this area[11, 15, 2, 3, 9, 28, 21]. Of late there has been a revived interest in language identification, in part due to the availability of a multi-language corpus[19] providing the means for comparative evaluations of techniques. Some proposed techniques for language identification combine feature vectors (filter bank, LPC, cepstum, formants) with prosodic features using polynomial classifiers[2], vector quantization[3, 9, 28], or neural nets[20]. Broad phonetic labels were used with finite state models[15] and with neural nets[20]. More recently, Gaussian mixture and HMM have been proposed for language identification[21, 31], as well as stochastic segment-based models[10].

This paper presents our recent work in language identification using phone-based acoustic likelihoods[5, 13]. The basic idea is to process in parallel the unknown incoming speech by different sets of phone models (each set is a large ergodic HMM) for each of the languages under consideration, and to choose the language associated with the model set providing the highest normalized likelihood.[1] Language identification can also be done using word recognition, but it is more efficient to use phone recognition, which has the added advantage of being task independent.

This approach has been evaluated for French/English language identification using laboratory quality speech, and for 10 languages using the OGI Multilingual telepone corpus[19]. Phone-based acoustic likelihoods have also been shown to be effective for sex and speaker-identification[5, 13]. In [18] it was found that the fine phonetic classes slightly outperformed broad phonetic categories, and both these outperformed acoustic features for Japanese-English language identification.

PHONE-BASED ACOUSTIC LIKELIHOODS

In this section we describe the use of phone-based acoustic likelihoods for the general case of identifying non-linguistic speech features such as language, gender, speaker, ... The basic idea is to train a set of large phone-based ergodic hidden Markov models (HMMs) for each non-linguistic feature to be identified. Feature identification on the incoming signal $\mathbf{x}$ is then performed by computing the acoustic likelihoods $f(\mathbf{x}|\lambda_i)$ for all the models λ_i of a given set. The feature value corresponding to the model with the highest likelihood is then hypothesized. This decoding procedure has been efficiently implemented by processing all the models in parallel using a time-synchronous beam search strategy.

This approach has the following characteristics:

- It can perform text-independent feature recognition. (Text-dependent feature recognition can also be performed.)

[1] In fact, this is not a new idea: House and Neuberg (1977)[11] proposed a similar approach for language identification using models of broad phonetic classes, where we use phone models. Their experimental results, however, were synthetic, based on phonetic transcriptions derived from texts.

- It is more precise than methods based on long-term statistics such as long term spectra, VQ codebooks, or probabilistic acoustic maps[27, 30].
- It can easily take advantage of phonotactic constraints.
- It can easily be integrated in recognizers which are based on phone models.

In our implementation, each large ergodic HMM is built from small left-to-right phonetic HMMs. The Viterbi algorithm is used to compute the joint likelihood $f(\mathbf{x}, \mathrm{s}|\lambda_i)$ of the incoming signal and the most likely state sequence instead of $f(\mathbf{x}|\lambda_i)$. This implementation is therefore a slightly modified phone recognizer with language-, sex-, or speaker-dependent model sets used in parallel, and where the output phone string is *ignored*[2] and only the acoustic likelihood for each model is taken into account.

The phone recognizer uses context-independent (CI) phone models, where each phone model is a 3-state left-to-right continuous density hidden Markov model (CDHMM) with Gaussian mixture observation densities. The covariance matrices of all Gaussian components are diagonal. Maximum likelihood estimators are used to derive language specific models whereas maximum a posteriori (MAP) estimators are used to generate sex- and speaker-specific models as has already been proposed in [7, 8].

In our original formulation, phonetic labels were required for training the models[5]. However, there is in theory no absolute need for phonetic labeling of the speech training data to estimate the HMM parameters. In this case, if a blind (or non informative) initialization for the HMM training re-estimation algorithm is used, the elementary left-to-right models are no longer related to the notion of phone. Such a non-informative initialization can lead to poor models for two reasons. First, the commonly used EM re-estimation procedure can only find a local maximum of the data likelihood and therefore "good" initialization is critical. Second, maximum likelihood training of large models with limited amount of training data (as in our case) cannot provide robust models if prior information information is not incorporated in the training process. We have experimented with two ways of dealing with these problems. The first is to use MAP estimation with seed models derived from transcribed speech data. We applied this approach to speaker identification in order to build the speaker-specific models from small amount of untranscribed speaker-specific data. The second approach is simply based on ML estimation where models trained on labeled data are used to generate an approximate transcription of the training data. We applied this second approach to language identification allowing us to estimate "phone" models from language specific data using a common phone alphabet for all of the languages. While there are many ways to introduce prior knowledge in the training process, it should be clear that the use of a great deal of prior information in the training procedure leads to more discriminative models.

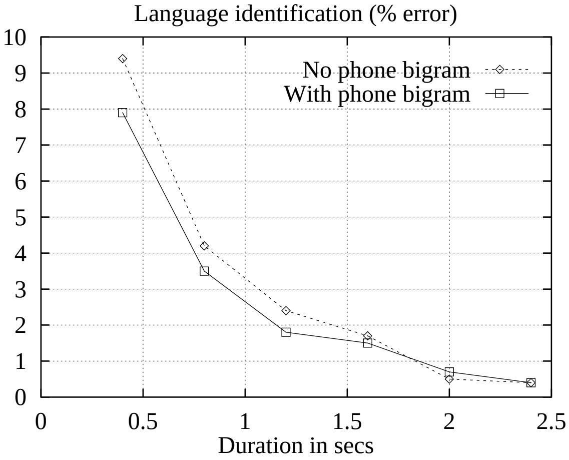

Figure 1: Overall French/English language identification as a function of duration with and without phonotactic constraints provided by a phone bigram. (The duration includes 100ms of silence.)

The use of ergodic HMM has been reported for speaker identification[24, 29, 16, 21] and for language identification[31] using small ergodic HMMs with a maximum of 5 to 8 states. Gaussian mixture models, which are special cases of ergodic HMM, have been used for speaker identification[25, 30]. The use of phone-based HMM has been reported for text-dependent[26, 17] and for text-independent, fixed-vocabulary[26] speaker identification.

FRENCH/ENGLISH LID EXPERIMENTS

Language-dependent models are trained using the BREF corpus for French and the WSJ0 corpus for English, containing read newspaper texts and similar size vocabularies[14, 23]. A set of 35 CI phone models were used for French and a set of 46 CI phone models for English. Each phone model has 32 gaussians per mixture, and no duration model is used. In order to minimize influences due to the use of different microphones and recording conditions a 4 kHz bandwidth was used. The training data for French include 2770 sentences from 57 speakers. For English the standard WSJ0 SI-84 training data (7240 sentences from 84 speakers) was used.

Corpus	*#sent.*	*0.4s*	*0.8s*	*1.2s*	*1.6s*	*2.0s*	*2.4s*
WSJ	100	5.0	3.0	1.0	2.0	1.0	1.0
TIMIT	192	9.4	5.7	2.6	2.1	0.5	0
BREF	130	8.5	1.5	0.8	0	0.8	0.8
BDSONS	121	7.4	2.5	2.5	1.7	0.8	0
Overall	543	7.9	3.5	1.8	1.5	0.7	0.4

Table 1: Language identification error rates as a function of duration and language (with phonotactic constraints).

Language identification accuracies are given in Table 1

[2] The likelihood computation can in fact be simplified since there is no need to maintain the backtracking information necessary to know the recognized phone sequence.

with phonotactic constraints provided by phone bigrams. Results are given for 4 test corpora, WSJ[23] and TIMIT[4] for English, and BREF[6] and BDSONS[1] for French, as a function of the duration of the speech signal which includes approximately 100ms of silence. The initial and final silences were automatically removed based on HMM segmentation, so as to be able to compare language identification as a function of duration without biases due to long initial silences. While WSJ sentences are more easily identified as English for short durations, errors persist longer than for TIMIT. In contrast for French with 400ms of signal, BDSONS data is better identified than BREF, perhaps because the sentences are phonetically balanced. For longer durations, BREF is slightly better identified than BDSONS. The LID performance is seen not to degrade in the cross-corpus condition.

Figure 1 shows the overall language identification results as a function of speech signal duration both with and without the use of phonotactic constraints. Using phonotactic constraints is seen to improve language identification, particularly for short signals. The error rate with 2s of speech is less than 1% and with 1s of speech is about 2%. With 3s of speech, language identification is almost error free.

OGI 10-LANGUAGE EXPERIMENTS

Language identification over the telephone opens a wide range of potential applications. Cognizant of this, we have evaluated our approach on the OGI 10 language telephone-speech corpus[19]. The Oregon Graduate Institute Multi-language Telephone Speech Corpus[19] was designed to support research on automatic language identification, as well as multi-language speech recognition. The entire corpus contains data from 100 native speakers of each of 10 languages (English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese). The utterances have been verified and transcribed at a broad phonetic level. The training data consists of calls from 50 speakers of each language. There are a total of about 4650 sentences, corresponding to about 1 hour of speech for each language. The test data are taken from the spontaneous stories from the development test data as specified by NIST[22] and include about 18 signal files for each language. Since these stories tend to be quite long, they have been divided into chunks by NIST, with each chunk estimated to contain at least 10 seconds of speech.

The training data was first labeled using a set of speaker-independent, context-independent phone models. Language-specific models were then estimated using MLE with the these labels. Thus, in contrast to the French/English experiments where the phone transcriptions were used to train the speaker-independent models, language-specific training is done *without* the use of phone transcriptions. 10-way language identification results are shown in Table 2 as a function of signal duration. The overall 10-language identification rate is 59.7% with 10s of signal (including silence).

Duration	*#10s chunks*	*2s*	*6s*	*10s*
English	63	54	64	67
Farsi	61	64	61	66
French	72	58	65	67
German	63	44	48	54
Japanese	57	28	32	42
Korean	44	48	48	55
Mandarin	59	46	51	61
Spanish	54	32	52	56
Tamil	49	69	82	82
Vietnamese	53	42	49	47
Overall	575	48.7	55.1	59.7

Table 2: OGI language identification rates (%) as a function of test utterance duration (without phonotactic constraints) for "10s chunks".

	Language Identified									
Lang.	*E*	*Fa*	*Fr*	*G*	*J*	*K*	*M*	*S*	*T*	*V*
English	42	1	2	10			1	1	4	2
Farsi	1	40	4	7	1	1				7
French	7	4	48	9	1		1		2	
German	13		7	34			1	6		2
Japan.		3	13	1	24	1		3		12
Korean	3	1	6		1	24	2			7
Mand.	2	2	5	6	2		36	1	1	4
Spanish	3	6	5	8	1			30		1
Tamil	1			5					40	3
Vietnam.	5	7	6			2		3	5	25

Table 3: 10-language confusion matrix for OGI corpus, "10s chunks".

There is a wide variation in identification accuracy across languages, ranging from 42% for Japanese to 82% for Tamil. The results of the language identification test as summarized by NIST[22] show similar variations in identification rate across languages for the different systems.

Table 3 shows the confusions obtained in language identification for the 10s chunks. Some confusions are seen to be symmetric between languages, for example, English and German are most likely to be confused with each other and French and German are also frequently confused. In contrast, Japanese is seen to be identified as French or Vietnamese, but neither of these languages are identified as Japanese.

Two-way French/English language identification was evaluated on the OGI corpus so as to provide a measure of the degradation observed due to the use of spontaneous speech over the telephone. The results are given in Table 4. Language identification was 82% at 10s (79% on French and

Duration	*#10s chunks*	*2s*	*6s*	*10s*
English	63	76	83	84
French	72	76	79	79
Overall	135	76	81	82

Table 4: French/English language identification rates (%) on the OGI corpus as a function of test for "10s chunks".

84% for English) for the 135 10s-chunks. This can be compared to the results with the laboratory read speech, where French/English language identification is better than 99% with only 2s of speech.

SUMMARY

In this paper we have applied the technique of phone-based acoustic likelihoods to the problem of language identification. The basic idea is to train a set of large phone-based ergodic HMMs for each language and to identify the language as that associated with the model set having the highest acoustic likelihood. The decoding procedure is efficiently implemented by processing all the language-specific models in parallel using a time-synchronous beam search strategy. This technique has also been successfully applied to gender and speaker identification[13] and has other possible applications such as for dialect identification (including foreign accents), or identification of speech disfluencies.

If the language can be accurately identified, it simplifies using speech recognition for a variety of applications, from selecting the language in multilingual spoken language systems to selecting an appropriate operator, or aiding with emergency assistance.

We would like to emphasize that the results on the OGI data are preliminary results which have been obtained by simply adapting the signal processing to the conditions of telephone speech. Our approach for French/English identification took advantage of the associated phonetic transcriptions, whereas for the OGI data, training was performed *without* the use of transcriptions. Despite these conditions, our results compare favorably to previously published results on the same corpus[20, 31, 10].

REFERENCES

[1] R. Carré, R. Descout, M. Eskénazi, J. Mariani, M. Rossi, "The French language database: defining, planning, and recording a large database," *ICASSP-84*.

[2] D. Cimarusti, "Development of an Automatic Identification System of Spoken Languages: Phase I," *ICASSP-82*.

[3] J.T. Foil, "Language Identification Using Noisy Speech," *ICASSP-86*.

[4] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM" NTIS order number PB91-100354.

[5] J.L. Gauvain, L. Lamel, "Identification of Non-Linguistic Speech Features ," *Proc. ARPA Workshop on Human Language Technology*, Mar. 1993.

[6] J.L. Gauvain, L.F. Lamel, M. Eskénazi, "Design considerations & text selection for BREF, a large French read-speech corpus," *ICSLP-90*.

[7] J.L. Gauvain, C.H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," *Proc. DARPA Speech & Nat. Lang. Workshop,* Feb. 1991.

[8] J.L. Gauvain, C.H. Lee, "Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities," *Speech Communication*, **11**(2-3), 1992.

[9] F.J. Goodman, A.F. Martin, R.E. Wohlford, " Improved Automatic Language Identification in Noisy Speech," *ICASSP-89*.

[10] T.J. Hazan, V.W. Zue, "Automatic Language Identification using a Segment-Based Approach," *EUROSPEECH-93*.

[11] A.S. House, E.P. Neuburg, "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations," *JASA*, **62**(3).

[12] L. Lamel, J.L. Gauvain, "Cross-Lingual Experiments with Phone Recognition," *ICASSP-93*.

[13] L. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *EUROSPEECH-93*.

[14] L. Lamel, J.L. Gauvain, M. Eskénazi, "BREF, a Large Vocabulary Spoken Corpus for French," *EUROSPEECH-91*

[15] K.P. Li, T.J. Edwards, "Statistical Models for Automatic Language Identification ," *ICASSP-80*.

[16] T. Matsui, S. Furui, "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs," *ICASSP-92*.

[17] T. Matsui, S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition," *ICASSP-93*.

[18] Y. Muthusamy et al., "A Comparison of Approaches to Automatic Language Identification Using Telephone Speech," *EUROSPEECH-93*.

[19] Y. Muthusamy, R. Cole, B. Oshika, "The OGI Multi-Language Telephone Speech Corpus," *ICSLP-92*.

[20] Y. Muthusamy,R. Cole,"Automatic Segmentation&Identification of Ten Languages Using Telephone Speech," *ICSLP-92*.

[21] S. Nakagawa, Y. Ueda, T. Seino, "Speaker-independent, Text-independent Language Identification by HMM," *ICSLP-92*.

[22] D.S. Pallett, A.F. Martin, "Language Identification: testing Protocols and Evaluations Procedures," *Proc. Speech Research Symposium XIII*, Baltimore, MD, June 1993.

[23] D. Paul, J. Baker, "The Design for the Wall Street Journal-based CSR Corpus" *Proc. DARPA Speech & Nat. Lang. Workshop*, Feb. 1992

[24] A.B. Poritz, "Linear Predictive Hidden Markov Models and the Speech Signal," *ICASSP-82*.

[25] R.C. Rose and D.A. Reynolds, "Text Independent Speaker Identification using Automatic Acoustic Segmentation," *ICASSP-90*.

[26] A.E. Rosenberg,C.H. Lee,F.K. Soong, "Sub-Word Unit Talker Verification Using Hidden Markov Models," *ICASSP-90*.

[27] A.E. Rosenberg, F.K. Soong, "Recent Research in Automatic Speaker Recognition," Ch. 22 in *Advances in Speech Signal Processing*, (Eds. Furui, Sondhi), Marcel Dekker, NY, 1992.

[28] M. Sugiyama, "Automatic Language Recognition Using Acoustic Features," *ICASSP-91*.

[29] N.Z. Tishby, "On the Application of Mixture AR Hidden Markov Models to Text-Independent Speaker Recognition," *IEEE Trans. Sig. Proc.*, **39**,(3), March 1991.

[30] B. Tseng, F. Soong, A. Rosenberg, "Continuous Probabilistic Acoustic MAP for Speaker Recognition," *ICASSP-92*.

[31] M. Zissman, "Automatic Language Identification using Gaussian Mixture and Hidden Markov Models," *ICASSP-93*.