Language Identification Using Phone-based Acoustic Likelihoods

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

As speech recognition technology advances, so do the aims of system designers, and the prospects of potential applications. One of the main efforts underway in the community is the development of speaker-independent, task-independent large vocabulary speech recognizers that can easily be adapted to new tasks. While the focus has been on improving the performance of the speech recognizers, it is also of interest to be able to identify what we refer to as some of the "non-linguistic" speech features present in the acoustic signal. For example, it is possible to envision applications where the spoken query is to be recognized without prior knowledge of the language being spoken. This is the case for information centers in public places, such as train stations and airports, where the language may change from one user to the next. The ability to automatically identify the language being spoken, and to respond appropriately, is possible. 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.

This paper presents our recent work in language identification using phone-based acoustic likelihoods[5, 7]. The basic idea is to process the unknown utterance by language-dependent phone models, identifying the language to be that language associated with the phone model set having the highest likelihood. This approach has been evaluated for French/English language identification in laboratory conditions, and for 10 languages using the OGI Multilingual telepone corpus[2]. Phone-based acoustic likelihoods have also been shown to be effective for sex and speaker-identification[5, 7].

PHONE-BASED ACOUSTIC LIKELIHOODS

A set of large phone-based ergodic hidden Markov models (HMMs) are trained for each non-linguistic feature to be identified (language, gender, speaker, ...). 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 be 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.)
- It is more precise than methods based on long-term statistics such as long term spectra, VQ codebooks, or probabilistic acoustic maps[10, 11].
- It can easily take advantage of phonotactic constraints.
- It can easily be integrated in recognizers which are based on phone models as all the components already exist.

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 nothing more than a slightly modified phone recognizer with feature-dependent model sets used in parallel, and where the output phone string is *ignored*[1] and only the acoustic likelihood for each model is taken into account.

The phone recognizer can use either context-dependent or context-independent phone models, where each phone model is a 3-state left-to-right CDHMM with Gaussian mixture observation densities. The covariance matrices of all Gaussian components are diagonal. Duration is modeled with a gamma distribution per phone model. Maximum likelihood estimators are used to derive language specific models.

EXPERIMENTAL RESULTS WITH FRENCH/ENGLISH LID

Language-dependent models are trained from similar-style corpora, BREF for French and WSJ0 for English, containing read newspaper texts and similar size vocabularies[8, 9]. For each language a set of context-independent phone models were built, 35 for French and 46 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 with phonotactic constraints provided by a phone bigram. Results are given for 4 test corpora, WSJ and TIMIT for English, and BREF and BDSONS for French, as a function of the duration of the speech

[1] 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.

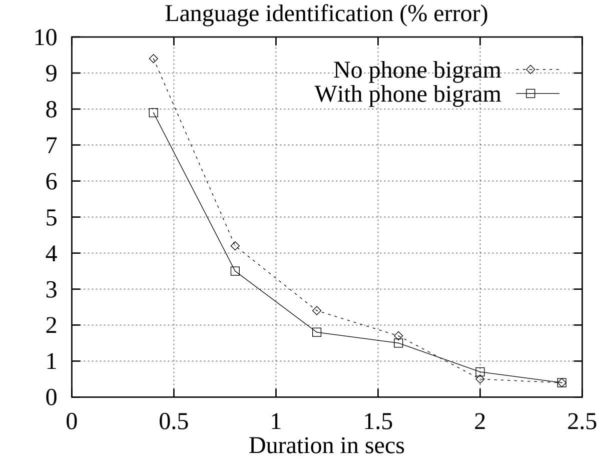

Figure 1: Overall French/English language identification as a function of duration with and without phonotactic constraints provided by a phone bigram. (The duration includes 100ms of silence.)

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 performance indicates that language identification is task independent.

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[2]. 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 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-specificic 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 59.4% with 10s of signal (including silence). There is a wide variation in identification accuracy across languages, ranging from 42% for Japanese to 82% for Tamil.

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".

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

Table 3: French/English language identification rates (%) on the OGI corpus as a function of test for "10s chunks".

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 3. Language identification was 82% at 10s (79% on French and 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.

We would like to emphasize that these are very preliminary results which have been obtained by simply porting the approach to the conditions of telephone speech. Our approach for English and French took advantage of the associated phonetic transcriptions, whereas for this evaluation the training has been performed *without* transcriptions. Despite these conditions, our results compare favorably to previously published results on the same corpus[3, 12].

REFERENCES

[1] R. Carré et al., "The French language database: defining, planning, and recording a large database," *ICASSP-84*

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

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

[4] J. Garofolo et al., "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 ," *ARPA Wshop Human Lang. Tech.*, Mar. 1993

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

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

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

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

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

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

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