Computer Speech and Language (1995) **9,** 87–103

A phone-based approach to non-linguistic speech feature identification

Lori F. Lamel and Jean-Luc Gauvain

LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

Abstract

In this paper we present a general approach to identifying non-linguistic speech features from the recorded signal using phone-based acoustic likelihoods. The basic idea is to process the unknown speech signal by feature-specific phone model sets in parallel, and to hypothesize the feature value associated with the model set having the highest likelihood. This technique is shown to be effective for text-independent gender, speaker and language identification. Text-independent speaker identification accuracies of 98·8% on TIMIT (168 speakers) and 99·2% on BREF (65 speakers), were obtained with one utterance per speaker, and 100% with two utterances for both corpora. Experiments in which speaker-specific models were estimated without using the phonetic transcriptions for the TIMIT speakers had the same identification accuracies as those obtained with the use of the transcriptions. French/English language identification is better than 99% with 2 s of read, laboratory speech. For spontaneous telephone speech from the OGI corpus, the language can be identified as French or English with 82% accuracy with 10 s of speech. The ten language identification rate using the OGI corpus was 59·7% with 10 s of signal.

1. Introduction

In this paper an approach for identifying non-linguistic speech features using phone-based acoustic likelihoods (Lamel & Gauvain, 1992, 1993*a,c*; Gauvain & Lamel, 1993) is presented. By the term non-linguistic we refer to speech features that give little information about the linguistic content of the message, such as the gender or identity of the speaker, the accent or even the language spoken. (Knowing the identity of the language is of little help in understanding what is being said if one does not already know the language.) While speaker and language identification have been the subject of long-term research, they have typically been seen as independent research areas, presenting different problems from those of speech recognition. In our view, these problems should be approached with the same modelling techniques as are used in speech recognition. For the most part, today's best performing speech recognition systems are based on a statistical modellization of the talker. From this point of view, message generation is represented by a language model which provides estimates of Pr(w) for all word strings w, and the acoustic channel encoding the message w in the

0885–2308/95/010087 + 17 $08.00/0

signal $\mathbf{x}$ is represented by a probability density function $f(\mathbf{x}|w)$. The speech decoding problem consists then of maximizing the *a posteriori* probability of w, or equivalently, maximizing the likelihood $\Pr(w)f(\mathbf{x}|w)$. The same modelling techniques can be adapted to other related applications, such as speech understanding and spoken language systems, or, as in this case, the identification non-linguistic speech features. For these, the feature decoding problem consists of maximizing the *a posteriori* probability of the feature λ (gender, speaker identity, etc.), or equivalently maximizing the likelihood $\Sigma_w \Pr(w|\lambda)f(\mathbf{x}|w, \lambda)$ if all the feature values are equally probable. The basic idea is to process the unknown speech signal by multiple feature-specific model sets in parallel (this is similar to the use of gender-dependent models for recognition), where instead of the output being the recognized string, the output is the characteristic associated with the model set having the highest likelihood. Some of these feature-specific models may also be directly used to more accurately model the speech signal thus in consequence improving the performance of the speech recognizer.

In the next section our approach using phone-based acoustic likelihoods is described. Section 3 provides a description of the corpora used in the experiments applying this approach to text-free identification of gender, speaker and language. Experimental results for these three problems are given in Sections 4, 5 and 6. A complete review of these problem areas is beyond the scope of this paper, in particular because excellent reviews already exist in the literature. Reviews of speaker identification and verification can be found in Atal (1976), Rosenberg (1976), Doddington (1985), Naik (1990) and Rosenberg and Soong (1992). Automatic language identification has also been the subject of long-term research (House & Neuberg, 1977; Li & Edwards, 1980; Cimarusti, 1982; Foil, 1986; Goodman, Martin & Wohlford, 1989; Sugiyama, 1991; Nakagawa, Ueda & Seino, 1992; Lamel & Gauvain, 1992; Gauvain & Lamel, 1993; Zissman, 1993). Recently gender identification has been of interest, primarily to improve acoustic modelling (Childers, Wu, Bae & Hicks, 1988; Fussell 1991; Gauvain, Lamel & Adda, 1993). The most closely related studies used small ergodic hidden Markov models (HMMs) for speaker identification (Poritz, 1982; Tishby, 1991; Matsui & Furui, 1992; Nakagawa *et al.*, 1992) or language identification (Zissman, 1993) and Gaussian mixture models (which are special case of ergodic HMMs) for speaker identification (Rose & Reynolds, 1990; Tseng, Soong & Rosenberg, 1992). The use of phone-based HMMs has been reported for text-dependent speaker identification or text-independent, fixed-vocabulary speaker identification (Rosenberg, Lee & Soong, 1990; Matsui & Furui, 1993). In this paper we demonstrate that all of these identification problems can be effectively handled by the use of phone-based acoustic likelihoods.

2. Phone-based acoustic likelihoods

A statistical modelling approach is taken, where the talker is viewed as a source of phones, modelled by a fully connected Markov chain. The lexical and syntactic structures of the language are approximated by local phonotactic constraints, and each phone is in turn modelled by a left-to-right HMM. This provides a better model of the talker than can be done with simpler techniques such as long-term spectra, VQ codebooks, or a simple Gaussian mixture. A set of phone-based HMMs is trained for each non-linguistic feature to be identified (language, gender, speaker, etc.). Feature identification on the incoming signal $\mathbf{x}$ is then performed by computing the phone-based 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 approach has the following characteristics:

- It can perform text-independent feature recognition. (Text-dependent feature recognition can also be performed.)
- By using a better model of the talker, it is more precise than methods based on long-term statistics such as long-term spectra, VQ codebooks, or probabilistic acoustic maps (Rosenberg & Soong, 1992; Tseng *et al.*, 1992).
- 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.

The Viterbi algorithm is used to compute the joint likelihood $f(\mathbf{x}, s|\lambda_i)$ 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 modified phone recognizer with language-, gender-, or speaker-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. This decoding procedure has been efficiently implemented by processing all the models in parallel using a time-synchronous beam search strategy.

The phone recognizer (Lamel & Gauvain, 1993*b*) can use either context-dependent or context-independent phone models, where each phone model is a three state left-to-right continuous density hidden Markov model (CDHMM) with Gaussian mixture observation densities. The covariance matrices of all Gaussian components are diagonal. Duration is modelled with a gamma distribution per phone model. As proposed by Rabiner, Juang, Levinson and Sondhi (1985), the HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search.

Maximum likelihood (ML) estimators are used to derive language-specific models whereas maximum *a posteriori* (MAP) estimators are used to generate gender- and speaker-specific models as proposed by Gauvain and Lee (1991, 1992*a*). The choice of method is closely linked to the amount of available training data and the task. ML is used to construct language-specific models since there is no reason to believe that each language should be derived from a common source of models and there is sufficient training data for each language. MAP estimation is used to construct gender-dependent models from a set of speaker-independent acoustic models, so as to be able to use all of the available data to train the largest model set possible, and also to take advantage of commonalities with the speaker-independent models. The speaker-independent seed models provide estimates of the parameters of the prior densities and also serve as an initial estimate for the segmental MAP algorithm (Gauvain and Lee, 1992*a*). This approach provides a way to incorporate prior information into the model training process and is particularly useful to build speaker-specific models when the amount of adaptation data is limited. More details of the specific acoustic modelling used for feature identification are given in Sections 4, 5 and 6, along with descriptions of the experiments.

In our original formulation, phonetic labels were required for training the models (Gauvain & Lamel, 1993). However, in theory there is no absolute need for phonetic labelling of the speech training data to estimate the HMM parameters. In this case, if

[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.

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 a limited amount of training data (as in our case) cannot provide robust models if prior 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 a small amount of untranscribed speaker-specific data. The second approach is simply based on ML estimation where models trained on labelled 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.

In the remainder of this paper experimental results applying our approach to text-free identification of gender, speaker and language are presented. In particular, we show that text-free identification of gender and speaker perform as well as fixed-text identification for a given duration of identification data, with the same quantity of training data.

3. Experimental conditions

In this section we provide a brief description of the corpora used to carry out experiments on identifying non-linguistic speech features, and provide a baseline performance assessment for the phone recognizer. Five corpora have been used: BDSONS and BREF for French; TIMIT and WSJ for English; and the OGI 10-language corpus. BREF, TIMIT and WSJ have been used for gender identification; BREF and TIMIT for speaker identification; and all five corpora have been used for language identification. Since the training and test data used differ for the various experiments, the details are specified for each experiment later.

3.1. The BDSONS corpus

BDSONS or Base de Données des Sons du Français (Carré, Descout, Eskénazi, Mariani & Rossi, 1984), was designed to provide a large corpus of French speech data for the study of the sounds in the French language and to aid speech research. The corpus contains an "evaluation" subcorpus consisting primarily of isolated and connected letters, digits and words from 32 speakers (16 male/16 female), and an "acoustic" subcorpus which includes phonetically balanced words and sentences. A subset of this latter subcorpus has been used for testing language identification.

3.2. The BREF corpus

BREF is a large read-speech corpus, containing over 100 h of speech material, from 120 speakers (55 male/65 female) (Lamel, Gauvain & Eskénazi, 1991). The text materials were selected verbatim from the French newspaper *Le Monde*, so as to provide a large vocabulary (over 20 000 words) and a wide range of phonetic environments (Gauvain, Lamel & Eskénazi, 1990). Containing 1115 distinct diphones and over 17 500 triphones, BREF can be used to train vocabulary-independent phonetic models. The text material was read without verbalized punctuation. All the data used for the experiments reported in this paper comes from the BREF80 subcorpus (two compact discs). Phonetic transcriptions of this subcorpus were automatically derived and manually verified (Gauvain & Lamel, 1992).

3.3. The DARPA TIMIT corpus

The DARPA TIMIT acoustic–phonetic continuous speech corpus (Garofolo, Lamel, Fisher, Fiscus, Pallett & Dahlgren, 1991) is a corpus of read speech designed to provide speech data for the acquisition of acoustic–phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from eight major dialect regions of the U.S.A. The TIMIT CDROM contains a training/test subdivision of the data that ensures that there is no overlap in the text materials. All of the utterances in TIMIT have associated time-aligned phonetic transcriptions.

3.4. The DARPA WSJ corpus

The DARPA Wall Street journal-based (WSJ) continuous speech recognition corpus (Paul & Baker, 1992) has been designed to provide general-purpose speech data (primarily read speech data) with large vocabularies. Text materials were selected to provide training and test data for 5 and 20 K word, closed and open vocabularies, and with both verbalized and non-verbalized punctuation. The recorded speech material supports both speaker-dependent and speaker-independent training and evaluation. In these experiments only data from the WSJ0 corpus were used. The standard WSJ0 SI-84 set of 7240 sentences from 84 speakers were used for training.

3.5. The 10-language OGI-TS corpus

The Oregon Graduate Institute multi-language telephone speech corpus (Musthusamy, Cole & Oshika, 1992) 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.

Since the identification of non-linguistic speech features is based on phone recognition, some baseline phone recognition results are given here for the corpora for which a phone transcription is available. The speaker-independent (SI) phone recognizers use sets of gender-dependent, context-dependent (CD) models which were automatically selected based on their frequencies in the training data which was used. Phone error rates with 428 CD models for BREF, 1619 for WSJ and 459 for TIMIT are given in Table I. For BREF and WSJ phone errors are reported after removing silences, whereas for TIMIT silences are included as transcribed, following the common practice for

TABLE I. Phone error (%) with context-dependent models and phonotactic constraints

Condition	No. of phones	Correct	Substitutions	Deletions	Insertions	Errors
BREF	35	81·7	13·7	4·6	3·0	21·3
WSJ nvp	46	79·3	16·2	4·5	5·0	25·7
TIMIT	39	78·3	16·7	4·9	4·9	26·6

TIMIT. The phone error for BREF is 21·3%, WSJ (Feb 92 5 knvp) is 25·7% and TIMIT (complete test set) is 26·6% scored using the 39 phone set proposed by Lee and Hon (1989). More details about the phone recognizer and experiments in phone recognition can be found in Lamel and Gauvain (1993*b*).

4. Gender identification

It is well known that the use of gender-dependent models gives improved word recognition performance over one set of speaker-independent models (Huang, Alleva, Hayamizu, Hon, Hwang & Lee, 1990). However, this approach can be costly in terms of computation for medium-to-large size tasks, since recognition of the unknown sentence is typically carried out twice—once for each gender. A logical alternative is to first determine the speaker's gender, and then to perform word recognition using the models of selected gender. Automatic identification of the speaker's gender has been previously investigated using single Gaussian classifiers (Childers *et al.*, 1988; Fussell, 1991), with gender identification accuracies reported for broad phonetic classes. Our approach is to use phone-based acoustic likelihoods for gender identification, using the same phone model sets that are used for phone or word recognition. The gender of the speaker is hypothesized as the gender associated with the model set giving the highest likelihood.

This approach was used in the LIMSI Nov 92 WSJ system (Gauvain *et al.*, 1993). The standard WSJ0 SI 84 training material, containing 7240 sentences from 84 speakers (42 male/42 female) was used to build speaker-independent CD phone models. Gender-dependent model sets were then obtained using MAP estimation (Gauvain & Lee, 1992b) with the SI seed models. The phone likelihoods using the CD male and female models were computed, and the gender of the speaker was selected as the gender associated with the model set that gave the highest likelihood. Since these male and female models are exactly the same CD phone models as used for word recognition, there is no need for additional training material or effort. No errors were observed in gender identification on the WSJ0 Feb 92 or Nov 92 5 k test data containing 851 sentences, from 18 speakers (10 male/8 female).

Gender identification was also assessed for French using a portion of the BREF corpus. Gender-dependent models were also obtained from SI seeds by MAP estimation. The training data consisted of 2770 sentences from 57 speakers (28 male/29 female). No errors in gender identification were observed on 109 test sentences from 21 test speakers (10 male/11 female).

To investigate gender identification on a larger set of speakers, the approach was evaluated on the 168 speakers of the TIMIT test corpus. SI seed models were trained

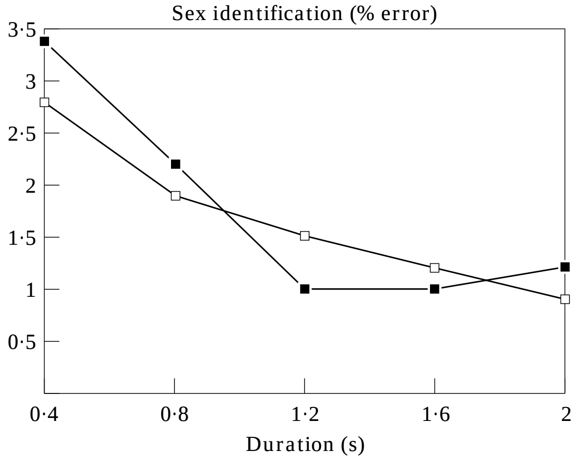

Figure 1. Text-independent (—□—) and text-dependent (—■—) gender identification error rates as a function of signal duration for 128 TIMIT speakers. (The duration includes 100 ms of silence.)

using all the available training data, i.e. 4620 sentences from 462 speakers. These models were then adapted using MAP estimation with data from the 326 male speakers and 136 female speakers to form gender-specific models. The test data consist of 1344 sentences, eight sentences from each of the 168 test speakers (112 male/56 female). The error rate in gender identification is shown as a function of the speech duration in Fig. 1. Each speech segment used for the test is part of a single sentence, and always starts at the beginning of the sentence, preceeded by about 100 ms of silence.[2] These results on this more significant test show that the text-independent gender identification error rate using phone-based acoustic likelihoods is 2·8% with 400 ms of speech and is about 1% with 2 s of speech. For reference, 400 ms of speech signal (which includes about 100 ms of silence) represents about four phones, i.e. the number found in a typical word in TIMIT (average 3·9 phones per word (Garofolo *et al.*, 1991)). This implies that before the speaker has finished enunciating the first word, one is fairly certain of the speaker's gender. We observed that the sentences misclassified with regards to the speaker's gender had better phone recognition accuracies with the cross-gender models.

An experiment of text-dependent gender identification was carried out using the same test data and the same phone models, in order to assess if by adding linguistic information the speaker's gender can be more easily identified. The basic idea was to measure the lower bound on the error rate that would be obtained if higher order knowledge such as lexical information were provided. To do this, a long left-to-right HMM was built for each gender by concatenating the gender-dependent CD phone models corresponding to the TIMIT transcription. The acoustic likelihoods were then computed for the two models. These likelihood values are evidently lower than are obtained for text-independent identification. The results are shown in the second curve of Fig. 1 where it can be seen that the error rate is not any better than the error rate

[2] The initial and final silences of each test sentence have been automatically reduced to 100 ms.

obtained with the text-independent method. This indicates that acoustic–phonetic information is sufficient to accomplish this task.

While in our previous work (Gauvain *et al.*, 1993), gender-identification was used primarily as a means to reduce the computation and to improve recognition performance, gender identification has other uses in spoken language systems. Accurate gender identification can permit the synthesis module of a system to respond appropriately to the unknown speaker. In languages like French, where the formalities are used more than in English, the system acceptance may be easier if greetings such as "Bonjour Madame" or "Je vous en prie Monsieur" are foreseen. Since gender identification is not perfect, some fall-back mechanism must be integrated to avoid including the signs of politeness if the system is unsure of the gender. This can be accomplished by comparing the likelihoods of the model sets, or by being wary of speakers for whom the better likelihood jumps back and forth between the gender-specific models over time.

5. Speaker identification

Speaker identification has been a topic of active research for many years (see for example, Atal, 1976; Rosenberg, 1976; Doddington, 1985; Naik, 1990; Rosenberg & Soong, 1992), and has many potential applications where propriety of information is a concern. In these experiments, the technique of phone-based acoustic likelihoods is applied to the problem of speaker-identification. A set of context-independent (CI) phone models was built for each speaker by adaptation of CI, SI seed models using MAP estimation (Gauvain & Lee, 1992*b*). The unknown speech was recognized by all of the speakers models in parallel, and the speaker identified as that associated with the model set having the highest likelihood. Speaker-identification experiments were performed using BREF for French and TIMIT for English. TIMIT has recently been used in a few studies on speaker identification (Rudasi & Zahorian, 1991; Bennani, 1992; Montacié & Le Floch, 1992) with high speaker identification rates reported using various sized subsets of the 630 speakers.

5.1. Experiments with BREF

For French, the acoustic seed models were 35 SI CI models, built using 2200 sentences from 57 BREF training speakers. Ten sentences for each speaker were reserved for adaptation and test. These models were adapted to each of 65 speakers (including eight new speakers not used in training the SI models) using eight sentences for adaptation. While the original CI models had a maximum of 32 Gaussians, the adapted models were limited to four mixture components, since the amount of adaptation data was relatively limited. The remaining two sentences were used for identification test. Text-independent speaker-identification results are given in the first entry in Table II for 65 speakers (27 male/38 female) as a function of signal duration. As for gender identification, the initial and final silences were adjusted to have a maximum duration of 100 ms according to the provided time-aligned transcriptions. Using only one sentence per speaker for identification, there is one error, corresponding to an identification accuracy of 99·2%. When two sentences for each speaker are used for identification test, all speakers are correctly identified.

Experiments for text-dependent speaker identification using exactly the same models

TABLE II. Text-independent vs. text-dependent speaker identification error rate as a function of duration for 65 speakers from BREF

Duration	0·5 s	1·0 s	1·5 s	2·0 s	2·5 s	EOS
BREF (text independent)	33·8	13·1	7·8	3·3	2·6	0·8
BREF (text dependent)	35·4	20·0	11·7	6·7	4·3	5·4

EOS is the end of sentence identification error rate. The duration includes 100 ms of silence.

and test sentences were performed. As can be seen in the second entry in Table II, the text-dependent error rates are higher than the text-independent error rates. There is almost a 4% degradation in the identification accuracy at the end of the sentence. These results were contrary to our expectations, in that typically text-dependent speaker verification is considered to outperform text-independent (Doddington, 1985; Rosenberg & Soong, 1992). However, Rosenberg and Soong (1992) have already demonstrated that with accurate modelling the difference in performance between text-dependent and text-independent speaker identification becomes quite small. A possible explanation of our results is that by using the phone transcription (i.e. text-dependent identification) the phone-based likelihoods are more dependent on the recognizer phone accuracy than for text-free identification. Therefore, speakers for whom the phone accuracies are lower than average, are more likely to be misidentified.

5.2. Experiments with TIMIT

For the experiments with TIMIT, a speaker-independent set of 40 CI models were built using data from all of the 462 training speakers. These SI CI models served as seed models to estimate 31 CI phone model sets for each of the 168 test speakers in TIMIT, using eight sentences (two SA, three SX and three SI) for adaptation. The remaining two SX sentences for each speaker were reserved for the identification test. This set of speakers was chosen for identification test so as to evaluate the performance for speakers not in the original SI training material, which greatly simplifies the enrollment procedure for new speakers. A reduced number of phones was used so as to minimize subtle distinctions, and to reduce the number of models to be adapted. As for BREF, while the original CI models had a maximum of 32 Gaussians, the adapted models were limited to four mixture components.

The 168 speaker-specific phone model sets were combined in parallel in one large HMM, which was used to recognize the unknown speech. Error rates are shown as a function of the speech signal duration in Fig. 2, for text-independent speaker identification. The curve labelled TIMIT-168 shows results with TIMIT SI seed models, using the phone transcription of the speaker-specific data during adaptation. The initial and final silences were adjusted to have a maximum duration of 100 ms according to the provided time-aligned transcriptions. If the entire utterance is used for identification, the accuracy is 98·5%. With 2·5 s of speech the speaker identification accuracy is 98·3%. For the small number of sentences longer than 3 s, identification is 100% correct, suggesting that if longer sentences were available performance would improve. This hypothesis is also supported by the result that speaker identification using both sentences

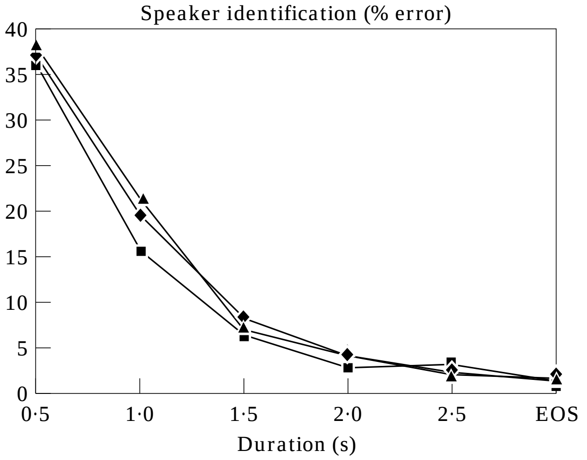

Figure 2. Text-independent speaker identification error rate as a function of duration for 168 test speakers of TIMIT. Training based on TIMIT seed models, with and without the phone transcription, and on WSJ seed models, without phone transcription. (EOS is the end of sentence identification error rate. The duration includes 100 ms of silence.) (—◆—), TIMIT-168 (TIMIT seed); (—▲—), TIMIT-168-NT (TIMIT seed); (—■—), TIMIT-168-NT (WSJ seed).

for identification is 100% correct. Text-dependent speaker identification on TIMIT exhibited the same performance degradation as observed for BREF. At the end of the sentence (EOS), the speaker-identification error is 6%, compared with 1·5% for text-independent identification with the same models.

Two additional experiments were performed in which speaker-specific models were estimated for each of the 168 test speakers in TIMIT without knowledge of the phonetic transcription. The same eight sentences were used for adaptation. In the first case, the 40 SI CI seed models from TIMIT were used to segment and label the data from the 168 speakers. In the second case, WSJ SI CI seed models were used to segment and label the TIMIT data. These labels were then used during the adaptation instead of the provided phone transcriptions. Performing text-independent speaker identification as before on the remaining two sentences gives the results shown in Fig. 2 TIMIT-168-NT. It can be seen that there is no significant difference in identification error when adaptation was performed with or without verified phone transcriptions, or when SI seed models from WSJ were used. The EOS identification error is 1·5% with TIMIT seed models and 1·2% with the WSJ seed models. As observed previously, if two sentences are used for identification, the speaker identification accuracy is 100%. This experimental result indicates that the time-consuming step of providing phonetic transcriptions is not needed for accurate text-independent speaker identification.

6. Language identification

While automatic language identification has been a research topic for over 20 years, there are relatively few studies published in this area. Of late there has been a revived interest in language identification, in part due to the availability of a multi-language corpus (Musthusamy *et al*., 1992) 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 (Cimarusti, 1982), vector quantization (Foil, 1986; Goodman *et al*., 1989; Sugiyama, 1991), or neural nets (Musthusamy & Cole, 1992). Broad phonetic labels were used with finite state models (Li & Edwards, 1980) and with neural nets (Musthusamy & Cole, 1992). More recently, Gaussian mixture and HMMs have been proposed for language identification (Nakagawa *et al*., 1992; Zissman, 1993).

Phone-based acoustic likelihoods can also be used for language identification. Once again, the basic idea is to process in parallel the unknown incoming speech by different sets of phone models (each set is a large HMM) for each of the languages under consideration, and to choose the language associated with the model set providing the highest normalized likelihood.[3] 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. Language identification can also be done using word recognition, but it is much more efficient to use phone recognition, which has the added advantage of being task independent.

6.1. *French/English LID experiments*

Experimental results for language identification for English/French were given in Lamel and Gauvain (1992, 1993*a*), where models trained on TIMIT and BREF were tested on different sentences taken from the same corpus. While these results gave high identification accuracies (100% if an entire sentence is used, and greater than 97% with 400 ms, and error free with 1·6 s of speech signal), it is difficult to discern that the language and not the corpus is being identified. Identification of independent data taken from the WSJ0 corpus was less accurate: 85% with 400 ms and 96% with 1·6 s of speech signal.

In these experiments we attempted to avoid the bias due to corpus, by testing both on other data from the same corpora from which the models were built, and on independent test data from different corpora. The language-dependent models are trained from similar style corpora, BREF for French and WSJ0 for English, both containing read newspaper texts and similar size vocabularies. A set of SI CI phone models were built for each language, with 35 models for French and 46 models for English.[4] Each phone model had 32 gaussians per mixture, and no duration model. In order to minimize influences due to the use of different microphones and recording conditions a 4 kHz bandwidth was used. The training data were the same as for gender identification (BREF: 2770 sentences from 57 speakers; WSJ0 SI-84: 7240 sentences from 84 speakers).

[3] In fact, this is not a new idea: House and Neuberg (1977) 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.

[4] The 35 phones used to represent French include 14 vowels (including three nasal vowels), 20 consonants (six plosives, six fricatives, three nasals and five semivowels), and silence. The phone table can be found in Gauvain and Lamel (1992). For English, the set of 46 phones include 21 vowels (including three diphthongs and three schwas), 24 consonants (six plosives, eight fricatives, two affricates, three nasals, five semivowels), and silence.

TABLE III. Language identification error rates as a function of duration and language with phonotactic constraints provided by a phone bigram

Test corpus	No. of sentences	Error rate vs. duration 0·4 s	0·8 s	1·2 s	1·6 s	2·0 s	2·4 s
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

The duration includes 100 ms of silence.

Language identification accuracies are given in Table III with phonotactic constraints provided by a phone bigram. Language identification error rates are given for the four test corpora, WSJ and TIMIT for English, and BREF and BDSONS for French, as a function of the duration of the speech signal. Approximately 100 ms of silence are included at the beginning and end of each utterance (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. The test data for WSJ consists of 100 sentences, the first 10 sentences for each of the 10 speakers (5 male/5 female) in the Feb 92-si 5 knvp (speaker-independent, 5 k, non-verbalized punctuation) test data. The TIMIT test data are the 192 sentences in the “coretest” set containing eight sentences from each of 24 speakers (16 male/8 female). The BREF test data consists of 130 sentences from 20 speakers (10 male/10 female) and for BDSONS the data is comprised of 121 sentences from 11 speakers (5 male/6 female).

While WSJ sentences are more easily identified as English for short durations, errors persist longer in these sentences than for TIMIT. In contrast for French, BDSONS data is better identified than BREF with 400 ms of signal, perhaps because the sentences are phonetically balanced. For longer durations, BREF is slightly better identified than BDSONS. Bearing in mind that the corpora were recorded under similar conditions, the performance demonstrated here shows that accurate task-independent, cross-corpus language identification can be achieved.

The overall French/English language identification error is shown in Fig. 3 as a function of duration, with and without phonotactic constraints provided by a phone bigram. Using the phone bigram is seen to improve language identification primarily for short signals. The overall error rate with 2 s of speech is less than 1% and with 1 s of speech (not shown) is about 2%. Incorporating phonotactic constraints had the smallest improvement for TIMIT, probably due to the nature of the selected sentences which emphasized rare phone sequences.

Language identification of the BREF and WSJ data is complicated by the inclusion of foreign words in the source text materials. One of the errors on BREF involved such a sentence. The sentence was identified as French at the beginning and then all of a sudden switched to English. The sentence was “Durant mon adolescence, je

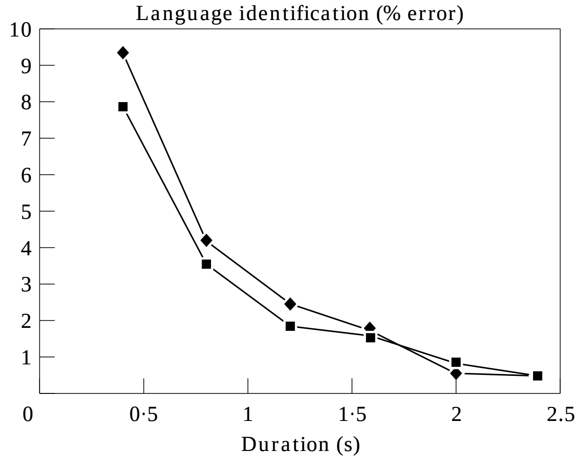

Figure 3. Overall French/English language identification as a function of duration with and without phonotactic constraints provided by a phone bigram. (The duration includes 100 ms of silence.) (—◆—), No phone bigram; (—■—), with phone bigram.

dévorais les récits westerns de *Zane Grey*, *Luke Short*, et *Max Brand* . . .", where the italicized words were pronounced in correct English.

We are in the process of obtaining corpora for other languages to extend this work. However, there are a variety of applications where a bilingual system, just French/English would be of use, including air traffic control (where both French and English are permitted languages for flights within France), telecommunications applications, and many automated information centres, ticket distributors and tellers, where already you can select between English and French with the keyboard or touch screen.

6.2. 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. The training data consists of calls from 50 speakers of each language. There are a total of about 4650 sentences, corresponding to about 1 h of speech for each language. The test data are taken from the spontaneous stories of the development test data as specified by National Institute of Standards and Technology 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 s of speech.

The training data for all languages was first labelled using a set of SI, CI phone models estimated on the NTIMIT corpus (Jankowski, Kalyanswamy, Basson & Spitz, 1990). Language-specific models were then estimated using ML estimators with these labels. Thus, in contrast to the French/English experiments where the phone transcriptions were used to train the SI models, language-specific training is done without the use of phone transcriptions. Language identification results using all 10 languages

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

Duration	No. of 10 s chunks	2 s	6 s	10 s
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 V. French/English language identification rates (%) on the OGI corpus as a function of test utterance duration for "10 s chunks"

Duration	No. of 10 s chunks	2 s	6 s	10 s
English	63	76	83	84
French	72	76	79	79
Overall	135	76	81	82

are shown in Table IV as a function of signal duration. The overall 10-language identification rate is 59·4% with 10 s of signal (including silence). There is a wide variation in identification accuracy across languages, ranging from 42% for Japanese to 82% for Tamil.

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 V. Language identification was 82% at 10 s (79% on French and 84% for English) for the 135 10 s chunks. This can be compared to the results with the laboratory read speech, where French/English language identification is better than 99% with only 2 s 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 favourably to previously published results on the same corpus (Musthusamy & Cole, 1992; Zissman, 1993).

7. Summary

In this paper we have presented an approach for the identification of non-linguistic speech features from recorded signals using phone-based acoustic likelihoods. We approach the problem with the same statistical modelling techniques as used in speech recognition. The basic idea is to train a set of phone-based HMMs for each non-linguistic feature to be identified (language, gender, speaker, etc.), and to identify the feature as that associated with the model having the highest acoustic likelihood of the set. While phone labels have been used to train the SI seed models, these models can then be used to label unknown speech, thus avoiding the costly process of transcribing the speech data. The decoding procedure is efficiently implemented by processing all the models in parallel using a time-synchronous beam search strategy.

This has been shown to be a powerful technique for gender, language and speaker identification, and has other possible applications such as for dialect identification (including foreign accents), or identification of speech disfluencies. Gender identification for BREF and WSJ was error-free, and 99% accurate for TIMIT with 2 s of speech. Speaker identification accuracies of 98·8% on TIMIT (168 speakers) and 99·1% on BREF (65 speakers) were obtained with one utterance per speaker, and 100% if two utterances were used for identification. This identification accuracy was obtained on the 168 test speakers of TIMIT without making use of the phonetic transcriptions during training, verifying that it is not necessary to have labelled adaptation data. SI models can be used to provide the labels used in building the speaker-specific models. Being independent of the spoken text, and requiring only a small amount of identification speech (in the order of 2·5 s), this technique is promising for a variety of applications, particularly those for which continual, transparent verification is preferable.

Tests of two-way language identification of read, laboratory speech show that with 2 s of speech the language is correctly identified as English or French with over 99% accuracy. Simply porting the approach to the conditions of telephone speech, the identification rate on the French and English data in the OGI multi-language telephone speech corpus was about 76% with 2 s of speech, and increased to 82% with 10 s of speech. The overall 10-language identification accuracy on the designated development test data of the OGI corpus is 59·7%. These results were obtained without the use of phone transcriptions for training, which had been used for the other experiments with laboratory speech.

References

Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE* **64**, 460–475.

Bennani, Y. (1992). Speaker identification through a modular connectionist architecture: evaluation on the TIMIT database. *Proceedings of the ICSLP-92*, Banff, Canada, pp. 607–610.

Carré, R., Descout, R., Eskénazi, M., Mariani, J. & Rossi, M. (1984). The French language database: defining, planning, and recording a large database. *Proceedings of the IEEE ICASSP-84*, San Diego, CA, pp. 1–4.

Cimarusti, D. (1982). Development of an automatic identification system of spoken languages: phase I. *Proceedings of the IEEE ICASSP-82*, Paris, France, **2**, 1661–1664.

Childers, D. G., Wu, K., Bae, K. S. & Hicks, D. M. (1988). Automatic recognition of gender by voice. *Proceedings of the IEEE ICASSP-88*, New York, NY, pp. 603–606.

Doddington, G. R. (1985). Speaker recognition—identifying people by their voices. *Proceedings of the IEEE* **73**, 1651–1664.

Foil, J. T. (1986). Language identification using noisy speech. *Proceedings of the IEEE ICASSP-86*, Tokyo, Japan, pp. 861–864.

Fussell, J. W. (1991). Automatic sex identification from short segments of speech. *Proceedings of the IEEE ICASSP-91*, Toronto, Canada, pp. 409–412.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. & Dahlgren, N. L. (1991). The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM. NTIS order number PB91-100354.

Gauvain, J. L. & Lamel, L. F. (1992). Speaker-independent phone recognition using BREF. *Proceedings of the DARPA Speech and Natural Language Workshop*, Arden House, NY, pp. 344–349.

Gauvain, J. L. & Lamel, L. F. (1993). Identification of non-linguistic speech features. *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, NJ, pp. 96–101.

Gauvain, J. L., Lamel, L. F. & Adda, G. (1993). LIMSI Nov 92 WSJ Evaluation. Presented at the DARPA Spoken Language Systems Technology Workshop, MIT, Cambridge, MA.

Gauvain, J. L., Lamel, L. F. & Eskénazi, M. (1990). Design considerations and text selection for BREF, a large French read-speech corpus. *Proceedings of the ICSLP-90*, Kobe, Japan, pp. 1097–1100.

Gauvain, J. L. & Lee, C. H. (1991). Bayesian learning of Gaussian mixture densities for hidden Markov models. *Proceedings of the DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, pp. 272–277.

Gauvain, J. L. & Lee, C. H. (1992*a*). MAP estimation of continuous density HMM: theory and applications. *Proceedings of the DARPA Speech and Natural Language Workshop*, Arden House, NY, pp. 185–190.

Gauvain, J. L. & Lee, C. H. (1992*b*). Bayesian learning for Hidden Markov model with Gaussian mixture state observation densities. *Speech Communication* **11**, 205–214.

Goodman, F. J., Martin, A. F. & Wohlford, R. E. (1989). Improved automatic language identification in noisy speech. *Proceedings of the IEEE ICASSP-89*, Glasgow, Scotland, pp. 528–531.

House, A. S. & Neuberg, E. P. (1977). Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *Journal of the Acoustical Society of America* **62**, 708–713.

Huang, X., Alleva, F., Hayamizu, S., Hon, H. W., Hwang, M. Y. & Lee, K. F. (1990). Improved hidden Markov modelling for speaker-independent continuous speech recognition. *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, pp. 327–331.

Jankowski, C., Kalyanswamy, A., Basson, S. & Spitz, J. (1990). NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. *Proceedings of the IEEE ICASSP-90*, Albuquerque, NM, pp. 109–112.

Lamel, L. F. & Gauvain, J. L. (1992). Continuous speech recognition at LIMSI. Final review of the Proceedings of the DARPA Artificial Neural Network Technology Speech Program, Stanford, CA.

Lamel, L. F. & Gauvain, J. L. (1993*a*). Cross-lingual experiments with phone recognition. *Proceedings of the IEEE ICASSP-93*, Minneapolis, MN, II, pp. 507–510.

Lamel, L. F. & Gauvain, J. L. (1993*b*). High performance speaker-independent phone recognition using CDHMM. *Proceedings of Eurospeech-93*, Berlin, Germany, I, pp. 121–124.

Lamel, L. F. & Gauvain, J. L. (1993*c*). Identifying non-linguistic speech features. *Proceedings of Eurospeech-93*, Berlin, Germany, I, pp. 23–28.

Lamel, L. F., Gauvain, J. L. & Eskénazi, M. (1991). BREF, a large vocabulary spoken corpus for French. *Proceedings of Eurospeech-91*, Genoa, Italy, pp. 505–508

Lee, K. F. & Hon, H. W. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on ASSP* **37**, 1641–1649.

Li, K. P. & Edwards, T. J. (1980). Statistical models for automatic language identification. *Proceedings of the IEEE ICASSP-80*, Denver, CO, pp. 884–887.

Matsui, T. & Furui, S. (1992). Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. *Proceedings of the IEEE ICASSP-92*, San Francisco, CA, II, 157–160.

Matsui, T. & Furui, S. (1993). Concatenated phoneme models for text-variable speaker recognition. *Proceedings of the IEEE ICASSP-93*, Minneapolis, MN, II, 391–394.

Montacié, C. & Le Floch, J. L. (1992). Ar-vector models for free-text speaker recognition. *Proceedings of the ICSLP-92*, Banff, Canada, pp. 611–614.

Musthusamy, Y. K. & Cole, R. A. (1992). Automatic segmentation and identification of ten languages using telephone speech. *Proceedings of the ICSLP-92*, Banff, Canada, pp. 1007–1010.

Musthusamy, Y. K., Cole, R. A. & Oshika, B. T. (1992). The OGI multi-language telephone speech corpus. *Proceedings of the ICSLP-92*, Banff, Canada, pp. 895–898.

Naik, J. M. (1990). Speaker verification: a tutorial. *IEEE Communications Magazine* **28**, 42–48.

Nakagawa, S., Ueda, Y. & Seino, T. (1992). Speaker-independent, text-independent language identification by HMM. *Proceedings of the ICSLP-92*, Banff, Canada, pp. 1011–1014.

Paul, D. & Baker, J. (1992). The design for the Wall Street journal-based CSR corpus. *Proceedings of the DARPA Speech and Natural Language Workshop*, Arden House, NY, pp. 357–362.

Poritz, A. B. (1992). Linear predictive hidden Markov models and the speech signal. *Proceedings of the IEEE ICASSP-82*, Paris, France, pp. 1291–1294.

Rabiner, L. R., Juang, B. H., Levinson, S. E. & Sondhi, M. M. (1985). Recognition of isolated digits using hidden Markov models with continuous mixture densities. *AT&T Technical Journal* **64**, pp. 1211–1234.

Rose, R. C. & Reynolds, D. A. (1990). Text independent speaker identification using automatic acoustic segmentation. *Proceedings of the IEEE ICASSP-90*, Albuquerque, NM, pp. 293–296.

Rosenberg, A. E. (1976). Automatic speaker verification: a review. *Proceedings of the IEEE* **64**, 475–487.

Rosenberg, A. E., Lee, C. H. & Soong, F. K. (1990). Sub-word unit talker verification using hidden Markov models. *Proceedings of the IEEE ICASSP-90,* Albuquerque, NM, pp. 269–272.

Rosenberg, A. E. & Soong, F. K. (1992). Recent research in automatic speaker recognition. In *Advances in Speech Signal Processing* (Furui & Sondhi, eds), Chapter 22, Marcel Dekker, New York.

Rudasi, L. & Zahorian, S. A. (1991). Text-independent talker identification with neural networks. *Proceedings of the IEEE ICASSP-91*, Toronto, Canada, pp. 389–392.

Sugiyama, M. (1991). Automatic language recognition using acoustic features. *Proceedings of the IEEE ICASSP-91*, Toronto, Canada, pp. 813–816.

Tishby, N. Z. (1991). On the application of mixture AR hidden Markov models to text-independent speaker recognition. *IEEE Transactions on Signal Processing* **39**, 563–570.

Tseng, B. L., Soong, F. K. & Rosenberg, A. E. (1992). Continuous probabilistic acoustic MAP for speaker recognition. *Proceedings of the IEEE ICASSP-92*, San Francisco, CA, II, 161–164.

Zissman, M. A. (1993). Automatic language identification using Gaussian mixture and hidden Markov models. *Proceedings of the IEEE ICASSP-93*, Minneapolis, MN, II, 399–402.

(*Received 25 February 1994 and accepted for publication 22 December 1994*)