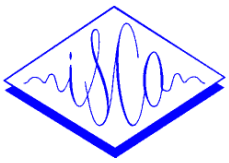

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

Identifying Non-Linguistic Speech Features

Lori F. Lamel and Jean-Luc Gauvain

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

10.21437/Eurospeech.1993-3

ABSTRACT

Over the last decade technological advances have been made which enable us to envision real-world applications of speech technologies. It is possible to foresee applications, for example, information centers in public places such as train stations and airports, where the spoken query is to be recognized without even prior knowledge of the language being spoken. Other applications may require accurate identification of the speaker for security reasons, including control of access to confidential information or for telephone-based transactions.

In this paper we present a unified 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 sex, speaker, and language identification and can enable better and more friendly human-machine interaction. 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 2 utterances for both corpora. Experiments estimating speaker-specific models *without* use of the phonetic transcription for the TIMIT speakers had the same identification accuracies obtained with the use of the transcriptions. French/English language identification is better than 99% with 2s of read, laboratory 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. 10 language identification using the OGI corpus is 59.7% with 10s of signal.

INTRODUCTION

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 specific tasks. It is becoming apparent that many of the portability issues may depend more on the specification of the task, and the ergonomy, than on the performance of the speech recognition component itself. The acceptance of speech technology in the world at large will depend on how well the technology can be integrated in systems which simplify the life of the users. This in turns means that the service provided by such a system must be easy to use, and as fast as other providers of the service (i.e., such as using a human operator).

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. If telephone-based applications are considered, a wide range of possibilities can be envisioned. These include emergency and medical assistance, travel services, communications related applications (translation services, operator and directory assistance, information services), as well as the well-known national intelligence applications.

Other applications of speech technology, such as for financial or banking transactions, access to confidential information, such as financial, medical or insurance records, etc., require accurate identification or verification of the user. Typically security is provided by the human who "recognizes" the voice of the client he is used to dealing with (the transaction often will also be confirmed by a fax), or for automated systems by the use of cards and/or codes, which must be provided in order to access the data. With the new payment and information retrieval services offered by telephone, it is a logical extension to explore the use of speech for user identification. An advantage of text-independent speaker verification techniques is that the speaker's identity can be continually verified during the transaction, in a manner completely transparent to the user. This can avoid the problems encountered by theft or duplication of cards, and pre-recording of the user's voice during an earlier transaction.

With these future views in mind, this paper presents a unified approach for identifying non-linguistic speech features using phone-based acoustic likelihoods. The basic idea is to process the unknown speech signal by multiple feature-specific phone model sets in parallel (this is similar to the use of sex-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.

A non-linguistic speech feature which has been the focus of many years of active research is the identity of the speaker. Reviews of speaker identification and verification can be found in [1, 39, 6, 33, 41]. Automatic language identification has also been the subject of long-term research[19, 26, 4, 7, 18, 43, 34, 21, 11, 46]. Recently sex-identification has been of interest, primarily to improve acoustic modeling capabilities[5, 8, 12]. We show that all of these identification problems can be effectively handled by the use of phone-based acoustic likelihoods.

PHONE-BASED ACOUSTIC LIKELIHOODS

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 (language, gender, speaker, ...). Feature identification on the incoming signal 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.)
- It is more precise than methods based on long-term statistics such as long term spectra, VQ codebooks, or probabilistic acoustic maps[41, 45].
- It can easily take advantage of phonotactic constraints. (These are shown to be useful for language identification.)
- 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 $f(\mathbf{x}, \mathbf{s}|\lambda_i)$ 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 language-, sex-, 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.

The phone recognizer can use either context-dependent or context-independent 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. Duration is modeled with a gamma distribution per phone model. As proposed by Rabiner et al.[37], the HMM and duration parameters are estimated separately and combined in the recognition process for the Viterbi search.

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 [15, 17]. The MAP estimates are obtained with the segmental MAP algorithm [16] using speaker-independent seed models. These seed models are used to estimate the parameters of the prior densities and to serve as an initial estimate for the segmental MAP algorithm. This approach provides a way to incorporate prior information into the model training process and is particularly useful to build the speaker-specific models when using only a small amount of speaker-specific data.

In our original formulation, phonetic labels were required for training the models[11]. 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.

The use of ergodic HMM has been reported for speaker identification[36, 44, 27, 34] and for language identification[46] 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[38, 45]. The use of phone-based HMM has been reported for text-dependent[40, 29] and for text-independent, fixed-vocabulary[40] speaker identification.

In the remainder of this paper experimental results applying our approach to text-free identification of sex, 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.

EXPERIMENTAL CONDITIONS

In this section we provide a brief description of the corpora used to carry out these experiments on identifying non-linguistic speech features, and provide a baseline performance assessment for the phone recognizer. Five corpora have been used in the experiments reported in this paper: BDSONS[3] and BREF[24, 14] for French; TIMIT[9] and WSJ[35] for English, and the OGI 10-language Corpus[32]. BREF, TIMIT and WSJ0 have been used for sex identification; BREF and TIMIT for speaker identification; and all 5 corpora have been used for language identification. Since the training and test data used differ for the various experiments, the details are specified later for each experiment.

The BDSONS Corpus: BDSONS, Base de Données des Sons du Français[3], was designed to provide a large corpus of

[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.

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 (16m/16f), and an "acoustic" subcorpus which includes phonetically balanced words and sentences. A subset of this latter subcorpus has been used for testing language identification.

The BREF Corpus: BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f)[24]. 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[14]. 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 sub-corpus (2 CDs). Phonetic transcriptions of this subcorpus were automatically derived and manually verified using a set of 35 phones[10].

DARPA TIMIT Corpus: The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus[9] 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 8 major dialect regions of the U.S. The TIMIT CDROM[9] 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.

DARPA WSJ Corpus: The DARPA Wall Street Journal-based Continuous-Speech Corpus (WSJ)[35] 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 5K and 20K 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 are used.

The 10-Language OGI-TS Corpus: The Oregon Graduate Institute Multi-language Telephone Speech Corpus[32] 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 we have a phone transcription. The speaker-independent (SI) phone recognizers use sets of sex-dependent, context-dependent (CD) models which were automatically selected based on their frequencies in the training data which was used. Phone errors rates with 428 CD models for BREF, 1619 for WSJ and 459 for TIMIT are given in Table 1. For BREF and WSJ phone errors are reported after removing silences, whereas for TIMIT silences are included as transcribed, following the common practise for TIMIT. The phone error for BREF is 21.3%, WSJ (Feb-92 5knvp) is 25.7% and TIMIT (complete test set) is 27.6% scored using the 39 phone set proposed by[25]. More details about the phone recognizer and experiments in phone recognition can be found in [23].

Condition	*#ph*	*Corr.*	*Subs.*	*Del.*	*Ins.*	*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

Table 1: Phone error (%) with CD models and phonotactic constraints.

SEX IDENTIFICATION

It is well known that the use of sex-dependent models gives improved word recognition performance over one set of speaker-independent models[20]. 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 sex. A logical alternative is to first determine the speaker's sex, and then to perform word recognition using the models of selected sex. Automatic identification of the speaker's sex has been previously investigated using single Gaussian classifiers[5, 8], with sex identification accuracies reported for broad phonetic classes. Our approach is to use phone-based acoustic likelihoods for sex-identification, using the same phone model sets that are used for phone or word recognition. The sex of the speaker is hypothesized as the sex associated with the model set giving the highest likelihood.

This approach was used in the LIMSI Nov-92 WSJ system[12]. The standard WSJ0 SI-84 training material, containing 7240 sentences from 84 speakers (42m/42f) was used to build speaker-independent CD phone models. Sex-dependent model sets were then obtained using MAP estimation[17] with the SI seed models. The phone likelihoods using the context-dependent male and female models were computed, and the sex of the speaker was selected as the sex 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 sex identification for WSJ0 on the Feb92 or Nov92 5k test data containing 851 sentences, from 18 speakers (10m/8f).

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

To investigate sex identification based on acoustic likelihoods on a larger set of speakers, the approach was evaluated on the 168 speakers of the TIMIT test corpus. SI seed models were trained using all the available training data, i.e., 4620 sentences from 462 speakers. These models were then adapted using

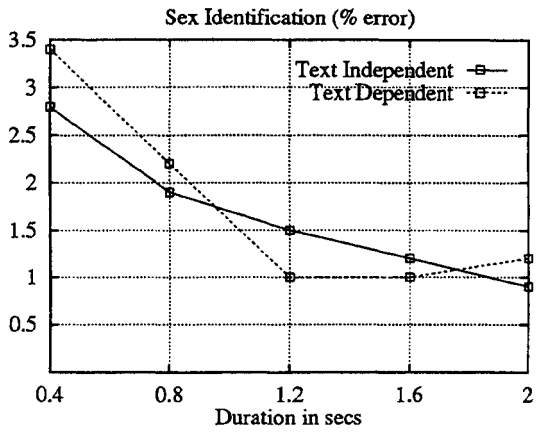

Figure 1: Text-independent and text-dependent sex-identification error rates as a function of signal duration for 128 TIMIT speakers. (The duration includes 100ms of silence.)

MAP estimation with data from the 326 males speakers and 136 females to form gender-specific models. The test data consist of 1344 sentences, 8 sentences from each of the 168 test speakers (112m/56f). The error rate in sex-identification is shown as a function of the speech duration in Figure 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 100ms of silence[2]. These results on this more significant test show that the text-independent sex identification error rate using phone-based acoustic likelihoods is 2.8% with 400ms of speech and is about 1% with 2s of speech. For reference, 400ms of speech signal (which includes about 100ms of silence) represents about 4 phones, i.e. the number found in a typical word in TIMIT (avg. 3.9 phones/word[9]). This implies that before the speaker has finished enunciating the first word, one is fairly certain of the speaker's sex. We observed that the sentences misclassified with regards to the speaker's sex had better phone recognition accuracies with the cross-sex models.

An experiment of *text-dependent* sex 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 sex by concatenating the sex-dependent CD phone models corresponding to the TIMIT transcription. The acoustic likelihoods were then computed for the two models. These likelihood values are lower than are obtained for text-independent identification. The results are shown in the second curve of Figure 1 where it can be seen that the error rate is not any better than the error rate obtained with the text-independent method. This indicates that acoustic-phonetic information is sufficient to accomplish this task.

While in our previous work[12], sex-identification was used primarily as a means to reduce the computation and to improve recognition performance, sex identification has other uses in spoken language systems. Accurate sex 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 sex-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 sex. 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.

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

SPEAKER IDENTIFICATION

Speaker identification has been a topic of active research for many years(see [1, 39, 6, 33, 41]), 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 CI phone models were built for each speaker by adaptation of CI, SI seed models using MAP estimation[17]. 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[42, 2, 30, 22] with high speaker identification rates reported using various sized subsets of the 630 speakers.

Experiments with BREF

For French, the acoustic seed models were 35 SI CI models, built using 2200 sentences from 57 BREF training speakers. 10 sentences for each were reserved for adaptation and test. These models were adapted to each of 65 speakers (including 8 new speakers not used in training the SI models) using 8 sentences for adaptation. While the original CI models had a maximum of 32 Gaussians, the adapted models were limited to 4 mixture components, since the amount of adaptation data was relatively limited. The remaining 2 sentences were used for identification test. Text-independent speaker-identification results are given in the first entry in Table 2 for 65 speakers (27m/38f) as a function of signal duration. As for sex identification, the initial and final silences were adjusted to have a maximum duration of 100ms 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 2 sentences for each speaker are used for identification test, all speakers are correctly identified.

Duration	*0.5s*	*1.0s*	*1.5s*	*2.0s*	*2.5s*	EOS
BREF (text ind.)	33.8	13.1	7.8	3.3	2.6	0.8
BREF (text dep.)	35.4	20.0	11.7	6.7	4.3	5.4

Table 2: Text-independent vs. text-dependent speaker identification error rate as a function of duration for 65 speakers from BREF. (EOS is End Of Sentence identification error rate. The duration includes 100ms of silence.)

Experiments for text-dependent speaker identification using exactly the same models and test sentences were performed. As can be seen in the second entry in Table 2, the text-dependent

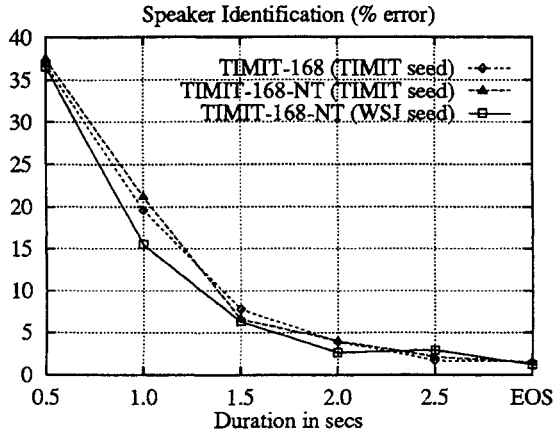

Figure 2: Text-independent speaker identification error rate as a function of duration for 168 test speakers of TIMIT. Training based on TIMIT seed models, with and without the phone transcription, and on WSJ seed models, without phone transcription. (EOS is End Of Sentence identification error rate. The duration includes 100ms of silence.)

error rates are higher error 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[6, 41]. However, Rosenberg et al. have already demonstrated that with accurate modeling the difference in performance between text-dependent and text-independent speaker identification becomes quite small[41]. 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.

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-phone model sets for each of the 168 test speakers in TIMI, using 8 sentences (2 SA, 3 SX, and 3 SI) for adaptation. The remaining 2 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 4 mixture components.

The 168 speaker-specific phone model sets were combined in parallel in one large HMM, which is used to recognize the unknown speech. Error rates are shown as a function of the speech signal duration in Figure 2, for text-independent speaker identification. The curve labeled 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 100ms according to the provided time-aligned transcriptions. If the entire utterance is used for identification, the accuracy is 98.5%. With 2.5s of speech the speaker identification accuracy is 98.3%. For the small number of sentences longer than 3s, identification was 100% correct, suggesting that if longer sentences were available performance would improve. This hyphothesis is also supported by the result that speaker-identification using both sentences for identification was 100% correct. ***Text-dependent speaker identification on TIMIT exhibited the same performance degradation as observed for BREF. At the EOS, the speaker-identification error is 6%, compared to 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 8 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 2 sentences gives the results shown in Figure 2 TIMIT-168-NT. It can be seen that there is not a significant difference in identification error when adaptation is performed with or without verified phone transcriptions, or when SI seed models from WSJ are used. The end of sentence identification error is 1.5% with TIMIT seed models and 1.2% with the WSJ seed models. As observed previously, if 2 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 accuracte text-independent speaker identification.

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[32] 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[4], vector quantization[7, 18, 43], or neural nets[31]. Broad phonetic labels were used with finite state models[26] and with neural nets[31]. More recently, Gaussian mixture and HMM have been proposed for language identification[34, 46].

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 ergodic 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,

[3] In fact, this is not a new idea: House and Neuberg (1977)[19] 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 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.

French/English LID Experiments

Experimental results for language identification for English/French were given in [21, 22], where models trained on TIMIT [9] and BREF [24], 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 400ms, and error free with 1.6s 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 400ms, and 4% error with 1.6s of speech signal.

In these experiments we attempted to avoid the bias due to corpus, by testing both on 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[14, 24, 35]. 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 has 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 is used. The training data were the same as for sex-identification (BREF:2770 sentences from 57 speakers and WSJ0 SI-84: 7240 sentences from 84 speakers).

Language identification accuracies are given in Table 3 with phonotactic constraints provided by a phone bigram. Language identification error rates are given for the 4 test corpora, WSJ and TIMIT for English, and BREF and BDSONS for French, as a function of the duration of the speech signal. Approximately 100ms 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 WSJ0 consist of 100 sentences, the first 10 sentences for each of the 10 speakers (5m/5f) in the Feb92-si5knvp (speaker-independent, 5k, non-verbalized punctuation) test data. For TIMIT, the 192 sentences in the "coretest" set containing 8 sentences from each of 24 speakers (16m/8f) was used. The BREF test data consists of 130 sentences from 20 speakers (10m/10f) and for BDSONS the data is comprised of 121 sentences from 11 speakers (5m/6f).

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 400ms 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.

[4]The 35 phones used to represent French include 14 vowels (including 3 nasal vowels), 20 consonants (6 plosives, 6 fricatives, 3 nasals, and 5 semivowels), and silence. The phone table can be found in [10]. For English, the set of 46 phones include 21 vowels (including 3 diphthongs and 3 schwas), 24 consonants (6 plosives, 8 fricatives, 2 affricates, 3 nasals, 5 semivowels), and silence.

Test Corpus	*# of sents*	*Error rate vs. Duration*					
		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 3: Language identification error rates as a function of duration and language with phonotactic constraints provided by a phone bigram. (The duration includes 100ms of silence.)

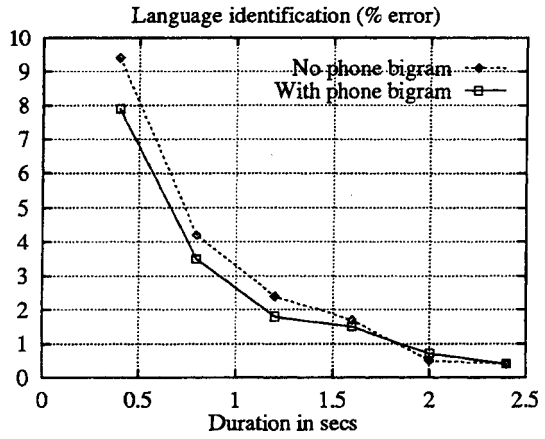

Figure 3: Overall French/English language identification as a function of duration with and without phonotactic constraints provided by a phone bigram. (The duration includes 100ms of silence.)

The overall French/English language identification error is shown in Figure 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 2s of speech is less than 1% and with 1s 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 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 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 centers, ticket distributors, and tellers, where already you can select between English and French with the keyboard or touch screen.

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[32]. 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.

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

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. Language identification results using all 10 languages are shown in Table 4 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	76	83	84
French	72	76	79	79
Overall	135	76	81	82

Table 5: 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 5. 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[31, 46].

SUMMARY

In this paper we have presented a unified approach for the identification of non-linguistic speech features from recorded signals using phone-based acoustic likelihoods. The inclusion of this technique in speech-based systems, can broaden the scope of applications of speech technologies, and lead to more user-friendly systems. The approach is based on training a set of large phone-based ergodic HMMs for each non-linguistic feature to be identified (language, gender, speaker, ...), and identifying the feature as that associated with the model having the highest acoustic likelihood of the set. 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 sex, language, and speaker-identification, and has other possible applications such as for dialect identification (including foreign accents), or identification of speech disfluencies. Sex-identification for BREF and WSJ was error-free, and 99% accurate for TIMIT with 2s 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 2 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 labeled data adaptation data. Speaker-independent 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 (on the order of 2.5s), 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 2s 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, French and English data in the OGI multi-language telephone speech corpus was about 76% with 2s of speech, and increased to 82% with 10s. The overall 10-language identification accuracy on the designated development test data of in the OGI corpus is 59.7%. These results were obtained *without* the use of phone transcriptions for training, which were used for the experiments with laboratory speech.

In conclusion, we propose a unified approach to identifying non-linguistic speech features from the recorded signal using phone-based acoustic likelihoods. This technique has been shown to be effective for text-independent, vocabulary-independent sex, speaker, and language identification. While

phone labels have been used to train the speaker-independent seed models, these models can then be used to label unknown speech, thus avoiding the costly process of transcribing the speech data. The ability to accurately identify non-linguistic speech features can lead to more performant spoken language systems enabling better and more friendly human machine interaction.

REFERENCES

[1] B.S. Atal, "Automatic Recognition of Speakers from Their Voices," *Proc. IEEE*, **64**,(4), April 1976.

[2] Y. Bennani, "Speaker Identification through a Modular Connectionist Architecture: Evaluation on the TIMIT Database," *ICSLP-92*.

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

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

[5] D.G. Childers, K. Wu, K.S. Bae, D.M. Hicks, "Automatic Recognition of Gender by Voice," *ICASSP-88*.

[6] G.R. Doddington, "Speaker Recognition - Identifying People by their Voices," *Proc. IEEE*, **73**,(11), Nov. 1985.

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

[8] J.W. Fussell, "Automatic Sex Identification from Short Segments of Speech," *ICASSP-91*.

[9] 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.

[10] J.L. Gauvain, L.F. Lamel, "Speaker-Independent Phone Recognition Using BREF," *Proc. DARPA Speech & Nat. Lang. Workshop*, Arden House, NY, Feb. 1992.

[11] J.L. Gauvain, L.F. Lamel, "Identification of Non-Linguistic Speech Features," *Proc. ARPA Human Language Technology Workshop*, Plainsboro, NJ, March 1993.

[12] J.L. Gauvain, L.F. Lamel, G. Adda, "LIMSI Nov92 WSJ Evaluation," presented at the *DARPA Spoken Language Systems Technology Workshop*, MIT, Cambridge, MA, Jan. 1993.

[13] J.L. Gauvain, L.F. Lamel, G. Adda, J. Mariani, "Speech-to-Text Conversion in French," to appear in *Int. J. Pat. Rec. & A.I.*, 1993.

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

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

[16] J.L. Gauvain, C.H. Lee, "MAP Estimation of Continuous Density HMM: Theory and Applications," *Proc. DARPA Speech & Nat. Lang. Workshop*, Arden House, NY, Feb. 1992.

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

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

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

[20] X. Huang, F. Alleva, S. Hayamizu, H.W. Hon, M.Y. Hwang, K.F. Lee, "Improved Hidden Markov Modeling for Speaker-Independent Continuous Speech Recognition," *Proc. DARPA Speech & Nat. Lang. Workshop*, Hidden Valley, PA, June 1990.

[21] L.F. Lamel, J.L. Gauvain, "Continuous Speech Recognition at LIMSI," Final review of the *Proc. DARPA Artificial Neural Network Technology Speech Program*, Stanford, CA, Sep. 1992.

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

[23] L.F. Lamel, J.L. Gauvain, "High Performance Speaker-Independent Phone Recognition Using CDHMM," *EUROSPEECH-93*.

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

[25] K.F. Lee, H.W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models," *IEEE Trans. ASSP*, 37(11), 1989.

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

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

[28] T. Matsui, S. Furui, "Speaker Recognition using Concatenated Phoneme Models," *ICSLP-92*.

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

[30] C. Montacié, J.L. Le Floch, "AR-Vector Models for Free-Text Speaker Recognition," *ICSLP-92*.

[31] Y.K. Musthusamy, R.A. Cole,"Automatic Segmentation and Identification of Ten Languages Using Telephone Speech," *ICSLP-92*.

[32] Y.K. Musthusamy, R.A. Cole, B.T. Oshika,"The OGI Multi- Language Telephone Speech Corpus," *ICSLP-92*.

[33] J.M. Naik, "Speaker Verification: A Tutorial," *IEEE Communications Magazine*, **28**(1), Jan. 1990.

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

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

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

[37] L.R. Rabiner, B.H. Juang, S.E. Levinson, M.M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *AT&T Technical Journal*, 64(6), 1985.

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

[39] A.E. Rosenberg, "Automatic Speaker Verification: A Review," *Proc. IEEE*, **64**,(4), April 1976.

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

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

[42] L. Rudasi, S.A. Zahorian, "Text-Independent Talker Identification with Neural Networks," *ICASSP-91*.

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

[44] 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.

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

[46] M.A. Zissman, "Automatic Language Identification Using Gaussian Mixture and Hidden markov Models," *ICASSP-93*.