Identification of Non-Linguistic Speech Features

Jean-Luc Gauvain and Lori F. Lamel

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

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 where the spoken query is to be recognized without even prior knowledge of the language being spoken, for example, information centers in public places such as train stations and airports. Other applications may require accurate identification of the speaker for security reasons, including control of access to confidential information or for telephone-based transactions. Ideally, the speaker's identity can be verified continually during the transaction, in a manner completely transparent to the user. With these views in mind, this paper presents a unified approach to identifying non-linguistic speech features from the recorded signal using phone-based acoustic likelihoods.

This technique is shown to be effective for text-independent language, sex, and speaker identification and can enable better and more friendly human-machine interaction. With 2s of speech, the language can be identified with better than 99% accuracy. Error in sex-identification is about 1% on a per-sentence basis, and speaker identification accuracies of 98.5% 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. An experiment using unsupervised adaptation for speaker identification on the 168 TIMIT speakers had the same identification accuracies obtained with supervised adaptation.

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 new 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.

Other applications, such as for financial or banking transactions, or 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 (and 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 widespread use of telephones, and 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 is that if text-independent speaker verification techniques are used, 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, such as the language being spoken, and the identity or sex of the speaker, using phone-based acoustic likelihoods. The basic idea is similar to that of using sex-dependent models for recognition, but instead of the output being the recognized string, the output is the characteristic associated with the model set having the highest likelihood. This approach has been evaluated for French/English language identification, and speaker and sex identification in both languages.

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 $\mathbf{x}$ is then performed by computing the acoustic likelihoods $f(\mathbf{x}|\lambda_i)$ for all the models λ_i of a given set. The feature value corresponding to the model with the highest likelihood is then hypothesized. This

decoding procedure can efficiently be implemented by processing all the models in parallel using a time-synchronous beam search strategy.

This approach has the following advantages:

- 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[26, 28].
- 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.

A disadvantage of the approach is that, at least in the current formulation, phonetic labels are required for training the models. However, there is in theory no absolute need for phonetic labeling of the speech training data to estimate the HMM parameters. Labeling of a small portion of the training data can be enough to bootstrap the training procedure and insure the phone-based nature of the resulting models. (In this case, phonotactic constraints must be obtained only from speech corpora.) We have sucessfully experimented with this approach for speaker identification.

In our implementation, each large ergodic HMM is built from small left-to-right phonetic HMMs. The Viterbi algorithm is used to compute the joint likelihood $f(\mathbf{x}, \mathrm{s}|\lambda_i)$ of the incoming signal and the most likely state sequence instead of $f(\mathbf{x}|\lambda_i)$. This implementation is therefore nothing more than a slightly modified phone recognizer with 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.[23], 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 [11]. The MAP estimates are obtained with the segmental MAP algorithm [16, 9, 10] 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 earlier reported results using this approach for language- and speaker-identification[13, 14, 7], the acoustic likelihoods were computed sequentially for each of the models. As mentioned earlier, the Viterbi decoder is now implemented as a one-pass beam search procedure applied on all the models in parallel, resulting in an efficient decoding procedure which saves a lot of computation.

EXPERIMENTAL CONDITIONS

Four corpora have been used to carry out the experiments reported in this paper: BDSONS[2] and BREF[15, 8] for French; and TIMIT[4] and WSJ0[22] for English. From the BDSONS corpus only the phonetically equilibrated sentence sub-corpus (CDROM 6) has been used for testing, whereas depending on experiment, the 3 other corpora have been used for training and testing.

The BDSONS Corpus: BDSONS, Base de Données des Sons du Français[2], 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 (16m/16f), and an "acoustic" subcorpus which includes phonetically balanced words and sentences from 12 speakers (6m/6f).

The BREF Corpus: BREF is a large read-speech corpus, containing over 100 hours of speech material, from 120 speakers (55m/65f)[15]. 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[8]. Containing 1115 distinct diphones and over 17,500 triphones, BREF can be used to train vocabulary-independenet phonetic models. The text material was read without verbalized punctuation.

The DARPA WSJ0 Corpus: The DARPA Wall Street Journal-based Continuous-Speech Corpus (WSJ)[22] 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.

The DARPA TIMIT Corpus: The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus[4] 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[4] contains a training/test subdivision of the data that ensures that there is no

[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.

overlap in the text materials. All of the utterances in TIMIT have associated time-aligned phonetic transcriptions.

Since the identification of non-linguistic speech features is based on phone recognition, some phone recognition results for the above corpora are given here. The speaker-independent (SI) phone recognizers use sets of context-dependent (CD) models which were automatically selected based on their frequencies in the training data. There are 428 sex-dependent CD models for BREF, 1619 for WSJ and 459 for TIMIT. Phone errors rates are given in Table 1. For BREF and WSJ phone errors are reported after removing silences, whereas for TIMIT silences are included as transcribed. Scoring without the sentence initial/final silence increases the phone error by about 1.5%. The phone error for BREF is 21.3%, WSJ (Feb-92 5knvp) is 25.7% and TIMIT (complete testset) is 27.6% scored using the 39 phone set proposed by[18]. These results are provided to calibrate the recognizers used in the experiments in this paper, and observe differences in the corpora. It appears that the BREF data is easiest to recognize at the phone level, and that TIMIT is more difficult than WSJ.

Condition	*Correct*	*Subs.*	*Del.*	*Ins.*	*Errors*
BREF	81.7	13.7	4.6	3.0	21.3
WSJ nvp	79.3	16.2	4.5	5.0	25.7
TIMIT	77.3	17.3	5.4	4.9	27.6

Table 1: Phone error (%) with CD models and phone bigram.

SEX IDENTIFICATION

It is well known that the use of sex-dependent models gives improved performance over one set of speaker-independent models. 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. This is the approach used in our Nov-92 WSJ system[6]. In these experiments the standard SI-84 training material, containing 7240 sentences from 84 speakers (42m/42f) is used to build speaker-independent phone models. Sex-dependent models are then obtained using MAP estimation[11] with the SI seed models. The phone likelihoods using context-dependent male and female models were computed, and the sex of the speaker was selected as the sex associated with the models that gave the highest likelihood. Since these CD male and female models are the same as are used for word recognition, there is no need for additional training material or effort. No errors were observed in sex identification for WSJ on the Feb92 or Nov92 5k test data containing 851 sentences, from 18 speakers (10m/8f).

For BREF, 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 further investigate sex identification based on acoustic likelihoods on a larger set of speakers, the approach was evaluted on the 168 speakers of the TIMIT test corpus. The SI seed models were trained using all the available training data, i.e., 4620 sentences from 462 speakers, and adapted using data from the 326 males speakers and 136 females to form gender-specific models. The test data consist of 1344 sentences, comprised of 8 sentences from each of the 168 test speakers (112m/56f). Results are shown in the first row of Table 2 where the error rate is given as a function of the speech duration. 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 sex identification error rate using phone-based acoustic likelihoods is 2.8% with 400ms of speech and is under 1% with 2s of speech. The 400ms of speech signal (which includes about 100ms of silence) represents about 4 phones, about the number found in a typical word (avg. 3.9 phones/word) in TIMIT. This implies that before the speaker has finished enunciating the first word, one is fairly certain of the speaker's sex. Sentences misclassified with regards to the speaker's sex had better phone recognition accuracies with the cross-sex models.

Using exactly the same test data and the same phone models, an experiment of text-dependent sex identification was carried out in order to assess if by adding linguistic information the speaker's gender can be more easily identified. To do this a long left-to-right HMM is built for each sex by concatenating the sex-dependent CD phone models corresponding to the TIMIT transcriptions. The basic idea is to measure the lower bound on the error rate that would be obtained if higher order knowledge such as lexical information were provided. The acoustic likelihoods are then computed for the two models. These likelihood values are lower than are obtained for text-independent identification. The results are given in the second row of Table 2 where it can be seen that the error rate is not any better than the error rate obtained with the text-independent method. This shows that acoustic-phonetic knowledge is sufficient to accomplish this task.

Duration	*0.4s*	*0.8s*	*1.2s*	*1.6s*	*2.0s*	*EOS*
Text indep.	2.8	1.9	1.5	1.2	0.9	1.2
Text dep.	3.4	2.2	1.0	1.0	1.2	1.3

Table 2: Error rate in sex identification as a function of duration. (EOS is End Of Sentence identification error rate.)

While in our previous work[6], sex-identification was used primarily as a means to reduce the computation, sex identification can permit the synthesis module of a system to respond appropriately to the unknown speaker. In French, where the

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

formalities are used perhaps more than in English, the system acceptance may be easier if the familiar "Bonjour Madame" or "Je vous en prie Monsieur" is 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 models.

LANGUAGE IDENTIFICATION

Language identification is another feature that can be identified using the same approach. In this case language-dependent models are used instead of sex-dependent ones. 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] In this way, it is no longer necessary to ask the speaker to select the language, before using the system. If the language can be accurately identified, it simplifies using speech recognition for a variety of applications, from 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.

Experimental results for language identification for English/French were given in [13, 14], where models trained on TIMIT [4] and BREF [15], 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 are 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 on data from the same corpora from which the models are 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[8, 15, 22]. For each language a set of context-independent phone models were built, 35 for French and 46 for English.[4] Each phone model has 32 gaussians per mixture, and no duration model is used. In order to minimize influences due to the use of different microphones and recording conditions a 4 kHz bandwidth is used. The training data were the same as for sex-identification on BREF (2770 sentences from 57 speakers) and WSJ (standard SI-84 training: 7240 sentences from 84 speakers).

Language identification accuracies are given in Tables 3 and 4 without and with phonotactic constraints provided by a phone bigram. Results are given for 4 test corpora, WSJ and TIMIT for English, and BREF and BDSONS for French, as a function of the duration of the speech signal which includes approximately 100ms of silence. As for speaker-identification, 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 are 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).

Duration	*0.4s*	*0.8s*	*1.2s*	*1.6s*	*2.0s*	*2.4s*
Eng. WSJ	7.0	3.0	2.0	2.0	1.0	1.0
Eng. TIMIT	10.9	6.3	3.1	2.1	0	0
Fr. BREF	10.8	2.3	2.3	0.8	0.8	0.8
Fr. BDSONS	7.5	4.1	1.7	1.7	0.8	0
Overall	9.4	4.2	2.4	1.7	0.5	0.4

Table 3: Language identification error rates as a function of duration and language (without phonotactic constraints).

Duration	*0.4s*	*0.8s*	*1.2s*	*1.6s*	*2.0s*	*2.4s*
Eng. WSJ	5.0	3.0	1.0	2.0	1.0	1.0
Eng. TIMIT	9.4	5.7	2.6	2.1	0.5	0
Fr. BREF	8.5	1.5	0.8	0	0.8	0.8
Fr. BDSONS	7.4	2.5	2.5	1.7	0.8	0
Overall	7.9	3.5	1.8	1.5	0.7	0.4

Table 4: Language identification error rates as a function of duration and language (with phonotactic constraints).

While WSJ sentences are more easily identified as English for short durations, errors persist longer than for TIMIT. In contrast for French with 400ms of signal, BDSONS data is better identified than BREF, perhaps because the sentences are phonetically balanced. For longer durations, BREF is slightly better identified than BDSONS. The performance indicates that language identification is task independent.

Using phonotactic constraints is seen to improve language identification, particularly for short signals. The smallest improvement is seen for TIMIT, probably due to the nature

[3] In fact, this is not a new idea: House and Neuberg (1977)[12] 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 3 nasal vowels), 20 consonants (6 plosives, 6 fricatives, 3 nasals, and 5 semivowels), and silence. The phone table can be found in [5]. 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.

of the selected sentences which emphasized rare phone sequences. The error rate with 2s of speech is less than 1% and with 1s of speech (not shown in the tables) is about 2%. With 3s of speech, language identification is almost error free.

Due to the source of the BREF and WSJ data, language identification is complicated by the inclusion of foreign words. 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 our language identification 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.

SPEAKER IDENTIFICATION

Speaker identification has been the topic of active research for many years (see, for example, [3, 21, 26]), and has many potential applications where propriety of information is a concern. In our experiments with speaker identification, a set of CI phone models were built for each speaker, by supervised adaptation of SI models[11], and the unknown speech was recognized by all of the speakers models in parallel.[5] 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[1, 20, 27, 14] with high speaker identification rates reported using subsets of 100 to all 462 speakers.

For the experiments with TIMIT, a speaker-independent set of 40 CI models were built using data from all of the 462 training speakers with 8kHz Mel frequency-based cepstral coefficients and their first order differences. 31-phone model sets were then adapted to each of the 168 test speakers using 8 sentences (2 SA, 3 SX, and 3 SI) for adaptation. We chose this set 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. The remaining 2 SX sentences for each speaker were reserved for the identification test. 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.

[5] Using HMM for speaker recognition has been previously proposed, see [26] for a review, and also [24, 25].

The unknown speech was recognized by all of the speakers models in parallel by building one large HMM. Error rates are shown as a function of the speech signal duration in Table 5, for text-independent speaker identification. As for sex and language identification, the initial and final silences were adjusted to have a maximum duration of 100ms according to the provided time-aligned transcriptions. Using the entire utterance the identification 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, speaker identification was correct, suggesting that with longer sentences performance will improve. This is also supported by the result that speaker-identification using both sentences for identification was 100%.

Duration	*0.5s*	*1.0s*	*1.5s*	*2.0s*	*2.5s*	EOS
TIMIT	36.9	19.6	7.8	3.9	1.7	1.5
BREF	33.8	13.1	7.8	3.3	2.6	0.8

Table 5: Text-independent speaker identification error rate as a function of duration for 168 test speakers of TIMIT, and 65 speakers from BREF. (EOS is End Of Sentence identification error rate.)

For French, the acoustic seed models were 35 SI CI models, built using data from 57 BREF training speakers, excluding 10 sentences to be used for adaptation and test. In order to have a similar situation to English, these models were adapted to each of 65 speakers (including 8 new speakers not used in training) using only 8 sentences for adaptation, and reserving 2 sentences for identification test. Using only one sentence per speaker for identification, there is one error, giving an identification accuracy of 99.2%, and when 2 sentences are used all speakers are correctly identified (as observed for TIMIT). Speaker-identification results are given in Table 5 for 65 speakers (27m/38f) as a function of signal duration. It can be noted that the identification accuracies as a function of time are similar for both corpora. However, since BREF sentences are somewhat longer than TIMIT sentences, the overall identification error rate per sentence is lower for BREF (EOS), even though the error for BREF at 2.5s is greater. For both TIMIT and BREF, when there was a confusion, the speaker was always identified by another speaker of the same sex.

Experiments for text-dependent speaker identification using exactly the same models and test sentences were performed. For both TIMIT and BREF a performance degradation was observed (on the order of 4% using the 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[3, 19].

An experiment was also performed in which speaker-adapted models were built for each of the 168 test speakers from TIMIT *without* knowledge of the phonetic transcription, using the same 8 sentences for adaptation. Performing text-independent speaker identification as before on the remaining 2 sentences give the results shown in Table 6. As be-

fore if both 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.

Duration	*0.5s*	*1.0s*	*1.5s*	*2.0s*	*2.5s*	EOS
TIMIT	37.5	21.2	6.6	4.0	2.1	1.5

Table 6: Text-independent speaker identification error rate as a function of duration for 168 test speakers of TIMIT with unsupervised adaptation. (EOS is End Of Sentence identification error rate.)

SUMMARY

In this paper we have reported on recent work on 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. With 2s of speech the language is correctly identified as English or French with over 99% accuracy. Speaker identification accuracies of 98.5% 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. The same identification accuracy was obtained on the 168 speakers of TIMIT using unsupervised adaptation, verifying that it is not necessary to provide phonetic transcription for accurate speaker identification. Being independent of the spoken text, and requiring only a small amount of speech (on the order of 2.5s), this technique is promising for a variety of applications, particularly those for which continual verification is preferable.

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 language, sex, and speaker identification and can enable better and more friendly human machine interaction.

REFERENCES

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

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

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

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

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

[6] 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.

[7] 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.

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

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

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

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

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

[13] L.F. Lamel, J.L. Gauvain, "Continuous Speech Recognition at LIMSI," *DARPA Speech & Nat. Lang. Workshop*, Sep. 1992.

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

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

[16] C.H. Lee, C.H. Lin, B.H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on ASSP*, April 1991.

[17] C.H. Lee, L.R. Rabiner, R. Pieraccini, J.G. Wilpon, "Acoustic modeling for large vocabulary speech recognition," *Computer Speech & Language*, 4, 1990.

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

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

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

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

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

[23] 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.

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

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

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

[27] M. Savic, J. Sorenson, "Phoneme Based Speaker Verification," *ICASSP-92*.

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