MULTILINGUAL PHONE RECOGNITION OF SPONTANEOUS TELEPHONE SPEECH

C. Corredor-Ardoy[*], *L. Lamel, M. Adda-Decker, J.L. Gauvain*

LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{corredor,lamel,madda,gauvain}@limsi.fr
http://www.limsi.fr/TLP

ABSTRACT

In this paper we report on experiments with phone recognition of spontaneous telephone speech. Phone recognizers were trained and assessed on IDEAL, a multilingual corpus containing telephone speech in French, British English, German and Castillan Spanish. We investigated the influence of the training material composition (size and linguistic content) on the recognition performance using context-independent Hidden Markov Models and phonotactic bigram models. We found that when testing on spontaneous speech data, using only spontaneous speech training data gave the highest phone accuracies for the four languages, even though this data comprises only 14% of the available training data. The use of context-dependent HMMs reduced the phone error across the 4 languages, with the average error reduced to 51.9% from the 57.4% obtained with CI models. We suggest a straightforward way of detecting non speech phenomena. The basic idea is to remove sequences of consonants between two silence labels from the recognized phone strings prior to scoring. This simple technique reduces the relative average phone error rate by 5.4%. The lowest phone error with CD models and filtering was obtained for Spanish (39.1%) with 4 language average being 49.1%.

INTRODUCTION

This paper presents some of our recent work in multilingual phone recognition of **spontaneous** telephone speech. Phone recognition has been the subject of long-term research at LIMSI[3, 6, 7, 8, 9]. We have previously reported on phone recognition experiments using high quality read speech (BREF[6, 7, 8, 9], TIMIT[7, 8, 9] and WSJ0[7, 8, 9]). We have also highlighted the importance of phonetic modeling, showing that the problems of language, speaker and sex identification can be addressed using a common phone-based acoustic likelihood approach[9]. However, very few phone recognition results have been reported on spontaneous speech corpora[3, 5]).

For the most part, speech recognition research is oriented at the word level, with no explicit evaluation at the phonetic level. In contrast, much of the research in language identification is focused on phonotactic modeling[11, 13] and determining measures for language scoring[13]. We believe that evaluating phone recognition is important for several reasons. First, phone accuracy has been shown to directly lead to improvements in word recognition accuracy[8] and in language identification[4]. Second, the analysis of phone recognition errors can be used to modify the lexicon (rectifying errors and including alternate pronunciations[8]) and even the phone set. Third, phone recognition may be used to assess *linguistic mismatch*. In the same way that acoustic mismatch may occur, linguistic mismatch exists when the training and test data differ in their linguistic contents. Speech corpora may include different types of data such as read isolated words or sentences, responses to precise or general questions, spontaneous monologues or conversations.

The experiments in this paper make use of the IDEAL corpus[3], a multilingual corpus containing telephone speech in French, British English, German and Castillan Spanish. The paper is organized as follows. First, the corpus and the evaluation protocol are described. Second, phone recognition experiments using context-independent (CI) Hidden Markov Models and backoff phonotactic bigram models are described. To address the problem of linguistic mismatch, we investigate phone recognition using acoustic models trained on different data subsets of the data and on all the available training data. We demonstrate that for recognition error of spontaneous telephone speech, acoustic models trained only on spontaneous speech outperform models trained on only read speech or on the entire corpus. For comparison we report phone recognition results using context-dependent (CD) Hidden Markov Models. Finally, we present a simple way of detecting and deleting non speech segments.

THE CORPORA

The IDEAL[3] corpus contains about 300 *matched calls* for each language (i.e., native French, British English, German and Castillan Spanish speakers calling from their home

[*] Corredor-Ardoy was with the LIMSI-CNRS when this work was carried out. He is now working at BOUYGUES TELECOM. 51, Avenue de l'Europe. 78944 Vélizy Cedex. France.

country) and up to 70 *crossed calls* for each language (i.e., native French speakers calling from the U.K., Germany and Spain, and native British English, German and Castillan Spanish speakers calling from within France). Each call covers a variety of data types: 12 direct questions to elicit responses, 18 items containing predefined texts to read and 6 questions aimed at collecting spontaneous monologues. We consider that the responses to the direct questions are linguistically closer to the read sentences than to the spontaneous speech. We have therefore established three training corpora using 250 calls for each language. The first subcorpus includes the responses to the direct questions and the read sentences (called the *read subcorpus*). The second subcorpus contains only the spontaneous speech (called the *spontaneous subcorpus*). The third subcorpus includes all the speech material (called the *read plus spontaneous corpus*). Phone error rates were assessed on 3 spontaneous monologues from at least 50 matched calls for each language. Table 1 shows the number of monologues for each language and subcorpus. On average there were 190 monologues per language, with more test calls for French and English. The amount of speech per monologue ranges from an average of 6.7s for French to 12.7s for Spanish.[1] The corpus was orthographically transcribed by native speakers of each language.

	Training (250 calls)			*Eval*
	[re]	[sp]	[re+sp]	[sp] (#calls)
French	7347	1170	8517	230 (77)
English	7457	1279	8736	217 (73)
German	7376	1594	8970	159 (53)
Spanish	7378	1204	8582	153 (51)
Average	7389	1312	8701	190

Table 1: Number of sentences for each language and corpus. [re]: read training corpus. [sp]: spontaneous training corpus. [re+sp]: read plus spontaneous training corpus (all training material). [sp]: evaluation corpus (spontaneous monologues).

THE EVALUATION PROTOCOL

All results are reported in terms of the phone error, based on a comparison between each *hypothesis* and *reference* phone string for the given monologue. By the term reference phone string, we refer to the phone transcription of the test sample which was obtained automatically by using the corresponding language-dependent phone recognizer to align the orthographic transcription with the acoustic signal. The recognizer automatically selects the most likely sequence of phones given the alternate pronunciations provided in the lexicon.[2] The hypothesis is the recognized phone sequence after removing silence labels. Phone error rates are calculated as the sum of the substitution, deletion and insertion errors divided by the number of phones in the reference string. Deletion and insertion penalties were applied to provide phone hypotheses with approximately equal length to the reference phone strings, so as to balance the deletion and insertion rates.[3]

EXPERIMENTS WITH CI MODELS

The first set of experiments were carried out using context-independent HMMs. Each phone was modeled by a three state continuous density HMM with 32 Gaussians per state. Sets of 35, 45, 48 and 25 phone units were used in French, British English, German and Castillan Spanish, respectively, where each set includes a language-dependent silence model.

In order to assess the impact of the linguistic mismatch on the acoustic models, different language-dependent phone recognizers were built using the three training corpora. Table 2 shows the phone recognition error rates on the spontaneous speech with phone recognizers trained on the read subcorpus ([re]), on the spontaneous subcorpus ([sp]) and on the read plus spontaneous corpus ([re+sp]). No phonotactic n-gram models were used. The lowest average phone error rate of 63.7% was obtained with acoustic models trained on the [sp] corpus, and the highest error of 69.4% was obtained with models trained on the [re] subcorpus. This effect is observed for each of the 4 languages. The phone error with the [sp] subcorpus is lower than that obtained with acoustic models trained on all the avaiable data (the [re+sp] corpus), even though this subcorpus contains only 14% of all available data (1.9 hours compared to 13.2 hours). This result illustrate the importance of having training data with is representative of the test data. Adding training data with different linguistical styles does not improve the performance of the phone recognizers.

	[re]{}	[sp]{}	[re+sp]{}
French	66.8%	**61.4%**	65.1%
English	76.0%	**73.6%**	74.8%
German	77.3%	**64.9%**	71.5%
Spanish	57.6%	**54.8%**	56.0%
Average	69.4%	**63.7%**	66.8%

Table 2: Phone error rates **on spontaneous speech** using CI HMMs. The columns correspond to language-dependent acoustic models trained on the read subcorpus ([re]), on the spontaneous subcorpus ([sp]) and on the read plus spontaneous corpus ([re+sp]). No phonotactic n-gram models were used ({}).

[1] There are 6.8s and 8.7s of speech for English and German respectively. The duration was obtained by summing the number of frames associated to the phone labels after forced alignment. This sum does not include silence frames. On average, the silence duration is about the same as the speech.

[2] In [8] we showed that the phone error rate is slightly lower using automatically obtained labels instead of manually generated references.

[3] It is possible to reduce the phone error by varying the insertion and deletion rates, however we have adopted the strategy of balancing these to more easily compare performances of different configurations.

In a second series of experiments we used phonotactic n-gram models with the best acoustic models (i.e., those estimated on the [sp] subcorpus). Different language-dependent phonotactic bigram backoff models were estimated on the three training corpora using the automatically generated phone transcriptions. Once again the best models correspond to training with the [sp] subcorpus: 57.4% compared to 60.0% and 62.3%, with backoff bigram models built on the [re+sp] and [re] corpora respectively, shown in Table 3).

	[sp]{re}	[sp]{sp}	[sp]{re+sp}
French	61.7%	**56.4%**	59.8%
English	71.4%	**67.2%**	69.4%
German	63.4%	**56.2%**	60.6%
Spanish	52.6%	**49.8%**	50.4%
Average	62.3%	**57.4%**	60.0%

Table 3: Phone error rates **on spontaneous speech** using context-independent HMMs. The columns give results with different bigram backoff models calculated on the read subcorpus ({re}), on the spontaneous subcorpus ({sp}) and on the entire ({re+sp}). The language-dependent acoustic models were trained on the spontaneous subcorpus ([sp])

These results appear to be related to the perplexity of the test data using the respective phonotactic models, as shown in Table 4. The perplexity of the test phone strings with a bigram backoff model estimated on the spontaneous subcorpus is 14.4 as compared to that with the read subcorpus (21.1) or the entire corpus (17.9). This difference in perplexity is consistent across the four languages.

	{re}	{sp}	{re+sp}
French	25.5	**15.8**	20.7
English	27.8	**17.8**	22.7
German	20.9	**13.5**	16.7
Spanish	13.2	**10.5**	11.6
Average	21.1	**14.4**	17.9

Table 4: Perplexities on the test monologues for the different bigrams backoff models: {re}, estimated on the read subcorpus ([re]); {sp} estimated on the spontaneous subcorpus ([sp]); and {re+sp}, estimated on the entire corpus ([re+sp]).

EXPERIMENTS WITH CD PHONE MODELS

Previous work with high quality read speech (BREF, TIMIT and WSJ0) has demonstrated that better phone accuracies can be obtained by using context-dependent (CD) HMMs[8]. The OGI_TS corpus[2] has been widely used in research related to phone recognition of spontaneous speech[1, 5, 9, 10, 11, 12]. To the best of our knowledge no results have been reported using this corpus with large sets of CD models.

The IDEAL corpus allows to estimate relatively large sets of context-dependent HMMs. We have built sets of 717, 862, 1011 and 917 CD phone models for French, British English, German and Castillan Spanish respectively. Each phone is a three state continuous density HMM with 32 Gaussians. The phone contexts to be modeled are based on their frequency of occurrence in the training data, with a mimimal count threshold of 100 occurrences. The models may be triphone models, right-context phone models, left context phone models or context-independent phone models. Table 5 gives results using the CD model sets on the spontaneous monologues with bigram backoff models estimated on the [sp] corpus.

	#models	*Corr*	*Sub*	*Del*	*Ins*	*Err*
French	717	61.0	25.6	13.4	11.4	**50.5**
English	862	52.3	33.4	14.4	13.1	**60.8**
German	1011	57.4	28.6	14.0	13.0	**55.6**
Spanish	917	70.7	16.9	12.4	11.6	**40.9**
Average	877	60.3	26.1	13.5	12.3	**51.9**

Table 5: Phone error rates **on spontaneous speech**, using context-dependent HMMs built on the [re+sp] corpus, with bigram backoff models calculated on the [sp] corpus. Number of CD phone models (#models CD). Correct (Corr), substitution (Sub), deletion (Del) and insertion (Ins) rates, and phone error (Err).

The average error rate of the 4 language-dependent phone recognizers is 51.9%. This corresponds to a 13.5% relative error reduction in as compared to the context-independent HMMs trained on the same data ([re+sp] corpus), and 9.6% of relative error reduction compared to the context-independent HMMs built on the [sp] corpus.

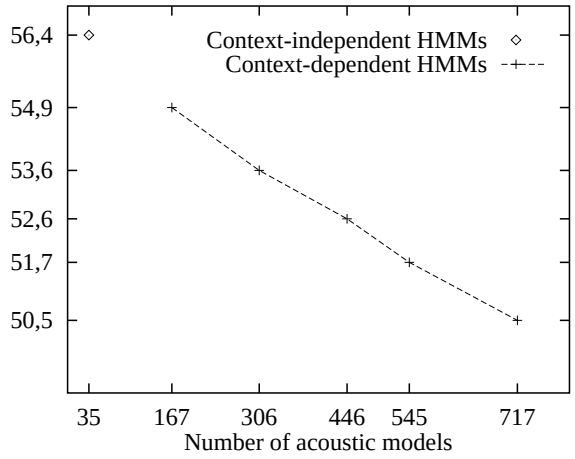

Figure 1: Phone error rates **on spontaneous speech** in French using differents sets of CD HMMs.

We have also evaluated the effect of varying the number of CD phone models. The experiments were carried out for French using different context-dependent HMM sets. Figure 1 gives the phone error rates on the spontaneous French data using 35 CI models and 167, 306, 446, 545 and 717 CD phone models. The lowest phone error rate was obtained using the highest number of CD phone models: 50.2% with 717 context-dependent HMMs.

DETECTING AND DELETING NON SPEECH PHENOMENA

A large number of the test monologues contain noises (microphone noise, tapping and other background noise or conversation) and non speech phenomena (breathing, cough, etc.), primarily at the beginning or end of the recording, or during long pauses. After analysis of the recognized phone strings, we have observed that such phenomena are often decoded by sequences of silence, plosives, fricatives and nasals sounds. An easy way to improve the noise robustness is to simply delete such phone sequences from the hypothesis strings prior to scoring. Based on this idea, we have developed language-dependent non speech filters. These filters remove consonant sequences found between two silence labels. Table 6 gives results using this technique on the hypothesized strings from Table 5. A 5.4% relative error reduction was obtained using this approach.

	#models	*Corr*	*Sub*	*Del*	*Ins*	*Err*
French	717	60.8	24.7	14.5	8.5	**47.7**
English	862	51.5	32.9	15.6	9.5	**57.9**
German	1011	57.0	28.0	14.9	8.7	**51.6**
Spanish	917	70.6	16.7	12.8	9.7	**39.1**
Average	877	60.0	25.5	14.5	9.1	**49.1**

Table 6: Phone recognition results **on spontaneous speech** using context-dependent HMMs built on the [re+sp] corpus and backoff bigram models calculated on the [sp] corpus. Non speech filters were applied to the hypothesis prior to scoring.

CONCLUSIONS

In this paper we have reported on experiments with multilingual phone recognition of spontaneous telephone speech. Phone recognizers were trained and assessed on the IDEAL corpus, containing multistyle speech in French, British English, German and Castillan Spanish. We evaluated the phone error using matched and mismatched linguistic styles, with context-independent HMM without phonotactic models. We observed that, for these conditions, acoustic models trained on all of the available training data did not have the best performance. In fact, when the training and test corpora contained the same type of data, the use of only a relatively small portion (1.9 h of the 13.2 h, or 14%) for training led to the best results: 63.7% versus 66.8% average error across the 4 languages). Adding more training data from different speech material appears to introduce linguistic mismatch which degrades the phone recognizer performance.

The use of a bigram backoff model estimated on the spontaneous subcorpus with acoustic models trained on the same data, provided the lowest error rate of 57.4%. These results may be explained by the perplexities: the best phonotactic models yield the lowest perplexities on the test corpus.

More accurate phone recognition rates were achieved using relatively sets of CD phone models (about 900), trained on the entire corpus. The average error rate across the 4 languages is 51.9%, corresponding to a relative error reduction of about 10% compared to CI models using the same phonotactic model.

We have also proposed a straightforward way to detect and delete non speech phenomena, which are frequent in spontaneous speech data. We observed that such events are often decoded by sequences of consonants surrounded by silences. Filtering the results of the context-dependent HMMs with the bigram backoff models yielded a relative error rate decreases of 5%. There is large range in performance, with lowest phone error rate of 39.1% was obtained for Castillan Spanish and the highest (59.7%) for British English. The superior performance on the Spanish data may be linked to the smaller set of phones used to describe the language, which enables more accurate acoustic and phonotactic modeling (lower entropy) for a given amount of training data. Accurate phone decoding of unconstrained spontaneous telephone speech data remains a challenging problem.

REFERENCES

[1] O. Andersen, P. Dalsgaard, "Language-identification based on Cross-Language Acoustic models and Optimised Information Combination," *Eurospeech'97*.

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

[3] C. Corredor-Ardoy, J.L. Gauvain, M. Adda–Decker, L. Lamel, "Language Identification With Language-Independent Acoustic Models," *Eurospeech'97*.

[4] C. Corredor-Ardoy, M. Adda-Decker, L. Lamel, J.L. Gauvain, "Identification Automatique de la Langue à travers le réseau téléphonique," Internal contract report CNET no. 7, Oct. 1997.

[5] J. Köhler, "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds," *ICSLP-96*.

[6] L. Lamel, J.L. Gauvain, "Experiments on Speaker-Independent Phone Recognition Using BREF," *ICASSP-92*.

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

[8] L. Lamel, J.L. Gauvain, "High Performance Speaker--Independent Phone Recognition Using CDHMM," *Eurospeech'93*.

[9] L. Lamel, J.L. Gauvain, "A phone-based approach to non-linguistic speech feature identification," *Computer Speech and Language*, **9**(1), Jan. 1995.

[10] T. Schultz, I. Rogina, A. Waibel, "LVCSR-Based Language-Identification," *ICASSP-96*.

[11] Y. Yan, E. Barnard, R.A. Cole, "Development of an approach to automatic language identification based on phone recognition," *Computer Speech and Language*, **10**(1), Jan. 1996.

[12] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. on SAP*, **4**(1), Jan. 1996.

[13] M.A. Zissman, "Predicting, Diagnosing and Improving Automatic Language Identification Performance," *Eurospeech'97*.