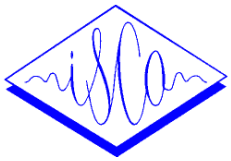

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

REAL-TIME MULTILINGUAL HMM TRAINING ROBUST TO CHANNEL VARIATIONS

E.E. Jan, Jaime Botella Ordinas, George Saon, and Salim Roukos

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598

ABSTRACT

This paper describes our efforts towards real-time telephony multi-lingual Large Vocabulary Continuous Speech Recognition (LVCSR) system. The trilingual (English, French and Spanish) landline cellular hybrid systems is compared to each of our best monolingual systems. The results are very comparable. The degradation is approximately less than 10%. A HMM state quality measurement technique is explored to improve the performances on multilingual acoustic models. A pilot experiment on English/Spanish bilingual system demonstrates very good results. We achieved between 5% to 20% improvement on different test conditions. To further extend to speaker phone applications, we employed different front-end processing techniques, mainly CDCN prior to HDA and MLLT to reduce the error rate on the trilingual system by as many as 30%. These results suggest that trilingual acoustic models can be used for real telephony applications.

1. INTRODUCTION

Advanced speech recognition technologies and affordable compute power has made spoken language man-machine interfaces increasingly attractive. Several consumer products are available on desktop dictation and command-and-control navigation. The argument for these conversational interfaces becomes even more compelling for "displayless" telephony applications such as directory assistance, voice-mail transcription, home banking, financial services for mutual fund and stock transaction, etc. Traditionally, telephony applications have been constrained to use cumbersome touch-tone interfaces requiring many keypresses to navigate complicated menu hierarchies. Speech input allows the telephone to become the ultimate thin client, providing ubiquitous conversational access to both traditional legacy IVR-based services and cutting edge web-derived content.

With the availability of speech recognition systems in multiple languages, it is interesting to explore the notion of universal speech recognizer for multiple languages with a universal phonology and a mechanism to handle language-dependent variations. In particular, regions like the New York metropolitan area or South California have bilingual populations conversant in both Spanish and English. Both English and French are used in Quebec and Montreal, Canada. This creates a potential problem for telephony applications. For example, in directory assistance, many speech recognition engines are run in a server, and the query may be requested by a different language at the same time. Furthermore, the query may contain mixed languages. Several schemes are available to solve this problem. A simple approach is employing two speech recognizers to decode each utterance. The outcome is chosen from the one with a better score. It is effective but rather expensive because the requirement of computation resources is doubled. Another approach is to employ a language identification system up front to choose the corresponding system. It can be cumbersome and will be more difficult to handle a mixed language query within a single utterance. In addition, the system load for different language severs can be unbalanced. This motivates the need for a seamless universal speech recognizer for multiple languages. Furthermore, we explore the possibility of a single acoustic models to handle landline, cellular and speakerphone channels.

2. ACOUSTIC MODELING ON TELEPHONY LVCSR SPANISH SYSTEM

Prior to exploring the telephony trilingual English/French/Spanish system, we need to establish the baselines: monolingual English, French and Spanish systems as well as bilingual English/French and English/Spanish systems. We have a great deal of context-rich Viavoice Spanish and French training data, which is recorded by close-talking microphone and many hours of telephony training data. However, this telephony data alone is insufficient and not context-rich to train a sophisticated telephony large vocabulary system. Here we use weighted multi-style training by pooling telephony training data with down-sampled close talking microphone data for system related to Spanish and French[6]. Also, To maintain unbiased to each individual language in the multilingual systems, equally amount of data from each language is used for training. Approximately 500 hours of training

Test set	English system	French System	Spanish System	English/French bilingual	English/Spanish bilingual	English/French/Spanish trilingual
English ATIS	6.1%	-	-	7.7%	6.9%	7.9%
French ATIS	-	16.4%	-	15.5%	-	17.0%
Spanish ATIS	-	-	6.7%	-	6.2%	6.6%

Table 1: Word error rates of ATIS test set on monolingual, bilingual and trilingual systems

data including landline and cellular is used for the trilingual system. To boost cellular phone performance, the cellular training data is weighted for multiple times. The system has approximately 2500 context dependent HMM with 45K Gaussian mixture. Additional IBM ViaVoice product level approximation is applied to make the system running under real-time.

In our approach, we do not to use language questions in the decision tree because: 1. small improvement is shown (5%) in ATIS domain training[2]. 2. More resources is required for training and testing. 3. We think, the language question is not very effective in trilingual and bilingual systems. Different languages have different phone sequences. This phone sequence is a distinguish feature for language identification, and has been widely used in language ID research. This language-specific phone sequence should have been separated in the decision tree without using the language questions.

2.1. Universal Phonology

From our experiment results show that by fixed the total number of parameters in range of 35 to 40 K Gaussian, approximately 10% relative improvement in performance can be achieve by reducing the number of Context Dependent HMM (CDHMM). This leads to a smaller phone set design. Those phones which are less distinguishable in telephony environment are merged. The new phone set contains 46, 34 and 38 phones for English, Spanish and French, respectively. These phones are then merged into a universal phone set. There are 67 phones in our universal phonology, of which 19 phones are shared in all three languages, 4 phones are shared in English and Spanish, 4 phones are shared by English and French, 17 phones, 11 phones and 11 phones are solely used by English, French and Spanish only, respectively. There are no phones shared by French and Spanish only.

2.2. ATIS Testset and ATIS Language Model

The ATIS (Air Travel Information System) domain test sets are first used in our studies. The reasons we choose this domain are: 1. This is a good telephony application. The perplexity and word error rate is reasonable to reflect the quality of the acoustic modeling. 2. It is easier to build the language models and can be extent to multilingual language models. 3. This application can be further extended to multilingual or cross lingual information retrieval studies. The speakers for Spanish testset come from all different places including Spain, Columbia, Mexico and Cube. The speakers for French testset also come from different places, e.g. France, Canada and USA. The Spanish and French test scripts are translated from English ATIS. These test utterances includes some US major city names which can degrade monolingual Spanish and French system performance.

Ten thousand English ATIS scripts and an equal amount of Spanish and French ATIS scripts, which were translated from English ATIS, were used to rebuild the trigram class language model. The class based trigram language models are built for monolingual and multi-lingual language systems. Major classes include city and airport names, time and airline companies.

2.3. Experimental results

The initial multi-lingual baseline on English, French and Spanish landline test set were approximately 10%, 20% and 15% respectively. After redesigning phone set, retraining the language model, rejecting bad training data, using mixing weights of multi-style training, and controlling the decision tree size, the error rates are reduced dramatically. The results (shown in Table 1) are compared to our best monolingual systems.

The word error rate of English ATIS on trilingual, bilingual English/French, bilingual English/Spanish and monolingual English system is 7.9%, 7.7%, 6.9% and 6.1%, respectively. The monolingual English system is trained on landline data only with much more training materials, however, the multilingual systems are landline/cellular hybrid system, this explains bigger degradation on English testset. The error rates of French and Spanish ATIS on these systems are very close. Several English cities are included in Spanish and French testset and the monolingual French and Spanish system are also landline cellular hybrid system, that explains better performance in bilingual systems. (15.5% vs. 16.4% in French ATIS, and 6.2% vs. 6.7% in Spanish) Unfortunately, we do not have language master for French during our work, the French ATIS language model is not as good as English and Spanish. This explains higher error

	N11.N28 44K	N12 34K	N20 57K
Stock500	14.2%	16.2%	15.2%
Stock2500	25.9%	28.7%	26.9%
Name500	2.5%	3.6%	3.2%
Name5000	9.2%	11.0%	9.2%

Table 2: Word error rates on different multilingual systems using HMM state quality measure.

rate on French ATIS testset. Overall, the trilingual system is very comparable to bilingual and monolingual system on this ATIS testset, and should be ready for real telephony applications.

3. ACOUSTIC MODELING USING HMM STATE QUALITY MEASUREMENT

Phones are shared in the multilingual system. Some of the HMM leaves are not well modeled and can be improved by using more parameters. A quality measurement of the HMM states are used to identify those confusible leaves. Advantage of the state quality measurement includes 1. Phones merged by expertise may not be aligned with current HMM technologies. Some of the shared phones can create more confusible context dependent leaves. 2. The language question is not used for building the decision tree. Some particular leaves include cross lingual contexts which maybe more confusible than regular context.

Several approaches have been proposed for state quality measurement. A common method is to decode the training data. The HMM state quality are then calculated by comparing the alignments between the decoded and reference scripts. However, the decoded alignment is the composite of acoustic model and language modeling score, hence the results can be biased by language models and search space of the decoding vocabulary. We use the rank based state quality measurement which is an approached solely rely on acoustic segments. The procedures are as followed. 1. For each frame in training data, find the top N ranked states. 2. Let R_i equal the number of frames that contain the correct state i in the top N state list. 3. Let C_i equal the total number of frames tagged as state i in the correct path. The rank based state quality measurement is defined as $P_i = \frac{R_i}{C_i}$

Clearly, this score represents the confusion of each state. By some thresholds, the HMM states can be clustered into different classes. Two classes, less confusible states and more confusible states, are used in our experiments.

Three bilingual English-Spanish systems are built with the same number of Context Dependent HMM (CDHMM) states and the results are presented using two English test sets, name and stock name testsets. Each test set are decoded using two flat grammars, a small and a large flat grammars. The flat grammars yield worse results than the weighted grammars. However, the accuracy is solely relied on detail match score, thus the performance is better matched to the quality of the acoustic models. The results are illustrated in Table 2. The first system, N11.N28, uses HMM state quality measurement to select the confusible states. Approximately 25% of the states are tagged as more confusible states, and 28 Gaussian mixtures are used. The others are models by 11 Gaussian mixtures. This system has total number of 44 K mixtures. The second, N12, and the third, N20, systems use 12 and 20 Gaussian mixture for each state, resulting in 34 K and 57 K Gaussian mixture, respectively. From the results, the N11.N28 system outperform N12 and N20 systems. Consistently, the N20 system always outperforms the N12 system. Therefore, the quality of multilingual acoustic model can be improvement using this rank based HMM state quality measurement.

4. TRILINGUAL SPEAKER PHONE MODELS

The channel variations for speakerphone are much bigger than landline and cellular. The variations are from speakerphone, and room acoustic, which varies dramatically even when the speech source location is changed. It is more difficult to have speaker phone training data to cover all the speaker phone conditions. Therefore, the multistyle training is less effective. Different front-end processing technique need to be used to improve the performance.

Linear Discriminant Analysis(LDA) is a very popular technique in front-end processing. The intent of LDA is to transform the feature space to a coordinate system in which useful information is concentrated in a smaller number of coordinates and where the coordinate values are uncorrelated. The latter condition is helpful if the pdf's are to be modeled by Gaussians with diagonal covariance matrices. LDA analysis, however, looks only at the global average of the within-class covariance matrices, and ignores differences between them. This leads the research to Heteroscedastic Discriminant Analysis (HDA) by removing the equal within-class covariance constraint. However, by removing this constraint, no close form solution is available for the object function. A numerical routine is required to calculate the projection matrix.

After applying the LDA or HDA, the covariance matrices for the Gaussian model is not diagonal. A Maximum Likelihood Linear Transformation (MLLT) by minimizing the loss from constraining the covariance matrices to be diagonal [5] can be applied then. Theoretically, better recognition accuracy can be achieved.

However, the above transformation techniques calculate the transformation matrix based on the training materials. The models are very biased to the training data, which is

	English	Trilingual system				
	CMN	CMN	HDA	HDA+MLLT	CDCN+HDA	CDCN+HDA+MLLT
Speaker Phone, Proper Name	31.0%	27.7%	26.0%	24.6%	23.0%	21.8%
Landline, Stock Name	14.0%	17.4%	13.1%	13.3%	12.8%	13.4%

Table 3: Word error rates of English Stock Name in landline condition and Proper Name in Speaker phone condition tested on trilingual systems with different signal processing. The results are also compared to English only system.

not preferred for speakerphone channel due to lack of training data. Our approach is using CDCN prior to the HDA and MLLT transformation. The CDCN algorithm has the advantage that it does not require a priori knowledge of the testing environment because it does not assume acoustic similarity between training and the testing data[8]. In general, the CDCN preprocessing transforms the input acoustic feature vectors into to a canonical space. The HDA and MLLT can then operate the transformation from the canonical space. The results will be less biased to training data and will be more robust to channel variation.

In our experiments, the HDA optimization is initialized with the LDA matrix. Every 9 consecutive 13 dimensional ceptral vectors are spliced together forming a 117 dimensional feature vectors. Multiple Gaussian mixtures per state are used to calculate the HDA object function. A full rank 39 dimensional MLLT is applied to transform the model into a diagonal covariance matrix.

The results of trilingual acoustic models with different front-end processing are reported in Table 3. The based line is a 39 dimensional Cepstrum Mean Normalization (CMN) trilingual system. The systems are further built with HDA and HDA plus MLLT. Finally, CDCN preprocessing are applied on both HDA only and HDA and MLLT systems.

The trilingual systems all have 2600 leaves with 45K Gaussian mixtures. These models are decoded used a speaker phone name test set and a landline stock name test set. The test set is decoded using flat grammars. The name grammars has 4000 entries, from Watson research center directory. This grammar contains significant amount of foreign name. The stock name grammar is a 10k flat grammars from US stock market. For comparison, the test set are also decoded on English monolingual system, which has 2500 leaves with 39 K Gaussian mixture.

From these results, as expected, the HDA and HDA+MLLT outperform the CMN system on speakerphone test set. However, additional huge gain (30% improvement over the CMN system) can be obtained by CDCN preprocessing.After CDCN preprocessing, all the data, including training data and speakerphone test set, is transformed to a canonical space. The HDA or HDA+MLLT can be more effective to improve the accuracy. The CMN trilingual system also outperform English monolingual system because part of the test data includes foreign name, which can be better presented via using universal phone set with the universal acoustic model. On the contrary, the stock name landline test set, monolingual English model outperform trilingual CMN model. The stock name are common English name and can be well modeled by English phone set. This landline stock test set is better matched to the channel characters of our training data. The HDA and HDA plus MLLT is sufficient enough to outperform CMN baseline. CDCN does not provide additional gain due to similar channel characteristic between training and testing. Fortunately, the CDCN does not cause significant performance degradation. We can then conclude that by combining CDCN with HDA (or HDA plus MLLT), the systems are more robust to channel variations.

5. REFERENCES

[1] K. Davies, et al, “The IBM Conversational Telephony System for financial applications”, Proceedings of Eurospeech 99, pp. 275-278, 1999

[2] T. Ward, et al, “Towards Speech Understanding across Multiple Languages”, Proceedings of ICSLP 98, pp 2243-2246, 1998

[3] L. R. Bahl, et al, “Robust methods for using context- dependent features and models in a continuous speech recognizer”, Proceedings of the ICASSP, pp 533-536, 1994.

[4] R. Lippmann, E. Martin, D. Paul, “Multi-style training for robust isolated-word recognition”, Proceedings of ICASSP97, pp. 705-708, 1987

[5] G. Saon, M. Padmanabhan, R. Gopinath and S. Chen “Maximum Likelihood Discriminant Feature Spaces”, proceedings of ICASSP2000, 2000

[6] E. Jan, J. Botella “Cross domains acoustic modeling”, proceedings of ICSLP2000, 2000

[7] Y. Gao, E. Jan, M. Padmanabhan, M. Picheny “HMM State Quality training”, Proceedings of ICASSP99, 1999

[8] F.H. Liu, A. Acero and R.M Stern, “Efficient Joint Compensation Of Speech For The Effects Of Additive Noise And Linear Filtering”, Proceedings. of ICASSP92, 1992 And Signal Processing, 1992.