Accent-Independent Universal HMM-Based Speech Recognizer for American, Australian and British English

Rathi Chengalvarayan

Lucent Speech Solutions
Lucent Technologies Inc.
2000 Lucent Lane, Naperville
Illinois 60566, USA
rathi@lucent.com

Abstract

This paper addresses the problem of speech recognition under accent variations in English language. It has been demonstrated in previous research efforts that the multi-transitional model architecture is one of the solutions for robust speech recognition. In this study, we describe an universal hybrid system that is trained with data from American, Australian, and British accented speech. Experimental results on connected-digit recognition task show an average string error rate reduction of about 62% and 8% when compared to our best monolingual and multi-transitional systems respectively. The result indicates that the universal model is about three times faster and half time smaller than the multi-transitional or multilingual models and this makes it an ideal choice for practical accent-independent speech recognition applications.

1. Introduction

As the demand for automatic speech recognition (ASR) technologies keeps on the rise, the need for the development of systems that are robust to speaking style, accents, environmental mismatch etc. is becoming increasingly important [11]. Speech recognition under accent variations is a challenging problem for which there are no completely satisfactory solutions [1]. This problem is crucial for the development of successful real-time multilingual applications in promising domains such as accent-independent speech recognition [14]. The speech for a particular language is rapidly changing depending on the regional accents [7]. Speech recognition suffers from significant performance deterioration when they are operated in mismatched accent conditions. Collecting data in an accent-dependent environment is a key factor to understanding and solving accent problems [4]. It has been demonstrated in previous research efforts, that the multi-HMMs and multi-transitional architectures are many of the proposed solutions for robust recognition [9]. The idea is to provide more variability to the system to be trained, and to support this variability with the greatest number of parameters. The main drawback is that the model size, and the computational complexity increases linearly related to different accents [3].

To alleviate the above problem we introduce an universal hybrid system that is trained with data recorded through Australian, American, and British accented speech for English language. This new universal system uses less than double the number of parameters as an individual system (American or Australian or British English) and significantly reduces the model parameters without affecting the performance when compared to multi-HMMs or multi-transitional models. We compared the performance of universal hybrid system on several independent test databases and demonstrated the effectiveness of a hybrid built with data taken from all three regional accented speech.

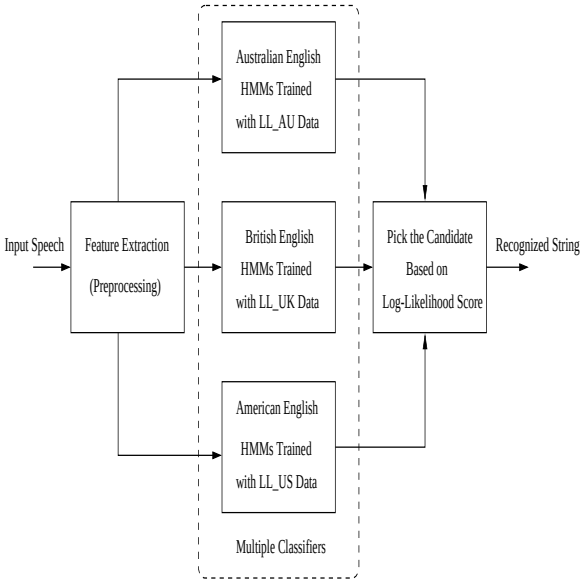

Figure 1: A block diagram of a typical parallel speech recognition system using multiple classifiers.

2. Universal Hybrid Systems

In this section the architecture of universal hybrid system is discussed. When the accent of a particular language is unknown, the important mismatch between training data and signal encountered in recognition phase decreases drastically the performances of the recognition systems [14]. Dialect also plays an important part in the overall degradation, resulting in different pronunciations for the same word [5]. There are many ways to reduce the accent and dialect variations within a given language [15]. A typical multiple-classifier approach is employing three accent-dependent speech recognizers to decode each utterance as shown in Fig. 1. The best hypothesis is chosen from the one with a top score. It is effective but rather expensive because the computation requirement is tripled [3]. Another approach is to

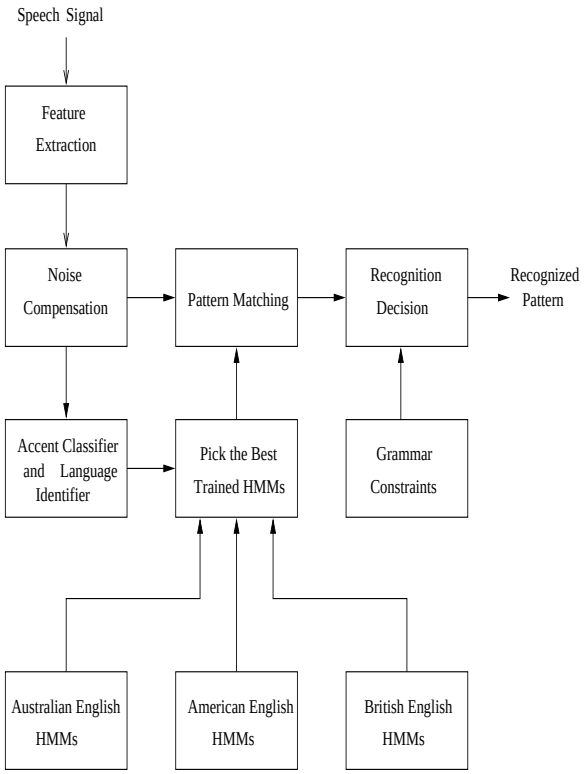

Figure 2: Structure of a speech recognition system based on accent classifier and language identifier.

integrate a accent classifier followed by a corresponding accent-specific recognizer as illustrated in Fig. 2. Many systems can get 100% correct accent classification when tested on training data, but can get an average of 81% on the 10 sec test utterance [1, 6, 15]. This can be cumbersome and will be difficult to handle more accent-specific utterances due to increased model complexity.

This motivates the need for simple accent-independent universal hybrid system (UNIV). This system uses a single decoder for British, Australian and American English digits, and is capable of recognizing digits with words from all accents as exemplified in Fig. 3. It is trained using a pooled data recorded through Australian, American, and British accented speech for English language. A negative side effect of this shared data is the increased possibility of confusion among words from three accents. This is overcome by doubling the Gaussian densities per state for the head and tails of a context-dependent head-body-tail topology. The reason for this is that those models are able to more adequately model the accent variations. Three monolingual context-dependent head-body-tail digit models for American (US), Australian (AU) and British (UK) English accents were trained using data from the corresponding accent [13]. Notice that the US model has 276 HMMs, the UK model has 304 HMMs and the AU model has 307 HMMs. The UK model has additional 28 context-dependent HMMs for the words *double* and *nought*. The AU model has three more HMMs for the word *triple* with silence contexts. For the purpose of comparison, multi-transitional (MULTI) models were created separately. Note that the multi-transition models were constructed by combining all the three monolingual models such that the decoder picks up the best model for a given

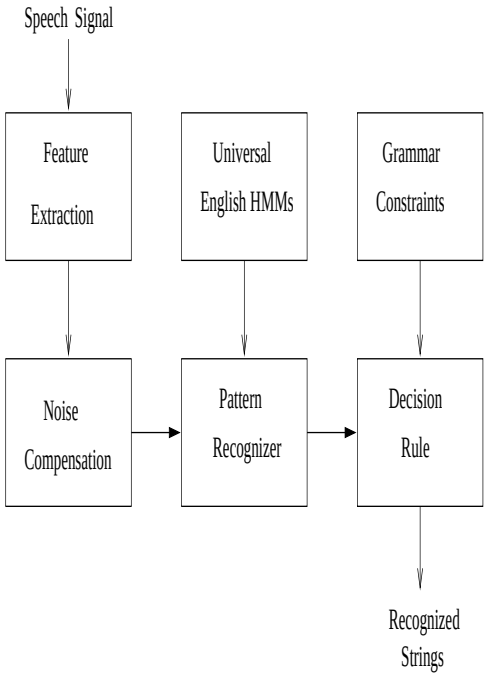

Figure 3: Block diagram of the HMM-based speech recognizer using universal modeling approach.

utterence from an unknown accent source. We call this model as *multiple pronunciation*, since each digit has three different pronunciation or accent variability [2].

3. Speech Database

This section describes the database, LL_US, used in this study [13]. This database is a good challenge for speech recognizers because of its diversity. It is a compilation of databases collected during several independent data collection efforts, field trials, and live service deployments. These independent databases are denoted as DB0 through DB3 and DB6. The LL_US database contains the English digits *one* through *nine*, *zero* and *oh*. It ranges in scope from one where talkers read prepared lists of digit strings to one where the customers actually use an recognition system to access information about their credit card accounts. The data were collected over network channels using a variety of telephone handsets. Digit string lengths range from 1 to 16 digits. The LL_US database is divided into two sets: training and testing. The training set, DB0 through DB3, includes both *read* and *spontaneous* digit input from a variety of network channels, microphones and dialect regions. The testing set is designed to have data strings from both matched and mismatched environmental conditions and includes all six databases. All recordings in the training and testing set are valid digit strings, totaling 7461 and 2023 strings for training and testing, respectively. The data distribution of the training and testing set is shown in Table 1. Only the 10, 14 and 16 digit strings were selected for testing.

The LL_UK consists of SpeechDat(M) database available through the European Lanuage Resource Association [16]. This database was collected over the U.K. landline telephone network. Recordings was done using an ISDN telephone interface, yielding 8 KHz, 8-bit samples A-law coded signals. Each cor-

Databases	Training		Testing	
	Strings	Speakers	Strings	Speakers
DB0	179	20	–	–
DB1	2568	500	–	–
DB2	2075	2075	518	518
DB3	2639	2639	713	713
DB6	–	–	792	1281
Total	7461	5234	2023	2512

Table 1: Regional distributions of spoken digit strings and the speaker population among the training and testing sets of the LL_US database.

Data Model	LL_AU	
	String Accuracy	Arc Count
AU	90.2%	16791
UK	75.4%	20453
US	46.2%	21701
MULTI	90.3%	81750
UNIV	90.1%	20717

Table 2: String accuracy and arc-count for a known-length connected-digit recognition task using landline Australian English (LL_AU) data as a function of various model type.

pus contains the speech of 1000 speakers (about 500 male and 500 female). Most items are read, some are spontaneously spoken. Speech material is conveniently split into two disjoint sets, a training one and a testing one. The LL_UK database contains the third pronounciation for zero as *nought*, and multiple related word such as *double*. The training database consists of digit string lengths range from 1 to 16 digits that were spoken by 700 speakers (350 male and 350 female) for a total of 2561 valid strings. The testing database has 300 speakers (107 male and 193 female) and only the valid digit strings were selected for a total of 505 strings. Only the digit strings with length of 4, 10, 11 and 16 were chosen for testing.

The LL_AU consists of SpeechDat(II) database that was collected over the Australian landline telephone network. This corpus contains the speech of 1000 speakers from all over the world. The training database consists of digit string lengths range from 1 to 16 digits that were spoken by 800 speakers for a total of 5298 valid strings. The testing database has 200 speakers and only the digit strings of length 5, 6, 10 and 16 were selected for a total of 848 strings. The LL_AU database contains the compact word *triple* in addition to LL_UK vocabulary. None of the speakers in the testing database appeared in the training databases.

4. Experimental Results

The recognizer feature set consists of 39 features that includes the 12 liftered linear predictive cepstral coefficients, log-energies, their first and second order derivatives [2]. The energy feature is batch normalized during training and testing [3]. Each feature vector is passed to the recognizer which models each word in the vocabulary by a set of left-to-right continuous mixture density HMM using context-dependent head-body-tail models. In this study, we model all possible inter-word coarticulation and each model is represented with 3 or 4 states, each having multiples of 4 mixture components. Silence is modeled with a single state model having 32 mixture components [2]. Training included updating all the parameters of the model, namely, means, variances and mixture gains using six epochs of MSE training [8]. Each training utterance is signal conditioned by applying batch-mode cepstral mean subtraction prior to being used in MSE training [3]. The number of competing string models was set to four, the step length was set to one and the length of the input digit strings are assumed to be unknown during the model training and a known-length grammar is used during testing. Penalties based on duration distributions are also applied to the likelihood score.

Data Model	LL_UK	
	String Accuracy	Arc Count
AU	81.2%	17160
UK	90.5%	16107
US	62.4%	18230
MULTI	89.5%	67885
UNIV	90.3%	18257

Table 3: String accuracy and arc-count for a known-length connected-digit recognition task using landline British English (LL_UK) data as a function of various model type.

We have conducted experiments to compare the performance of universal hybrid system on several independent test databases and to demonstrate the effectiveness of a hybrid built with data taken from all three regional accented speech. The Table 2 through 4 presents the string accuracy and arc counts for three different monolingual models using all the three datasets. When the LL_AU data is tested on a US system, the arc-count increases tremendously to a point where the recognizer is overloaded with unnecessary log-likelihood computations, arc-expansions, etc. which results in a longer delay in reporting the recognized string. This is true to some extent that the US models look more fuzzier when tested on mismatched LL_AU data and hence the average-arc-count gets incremented and slow down the decoding speed. Similarly when the LL_US data is tested using AU model the arc count increases due to mismatched testing data. The same observation can also be made for British accented data and the correponding UK model. The monolingual model runs faster and gives the best string accuracy when tested on the matched data. Notice that the less number of arcs relates to less computational complexity. The mismatch between AU model and LL_UK data is minimal when compared to AU model tested on US data. We can clearly see the mismatch in performance between the three different regional accents. Further test results using multiple classifier (MULTI) and universal hybrid model (UNIV) are tabulated in Table 2 through 4 for LL_UK, LL_AU and LL_US databases.

Table 5 shows the average string accuracy, model size and average arc counts for five different models. The US system yields the worst and gets about 67.4% and the AU system yields about 75.0% of string accuracy. The UK model is the best among the monolingual system and provides about 76.8% of string accuracy with little more average arc counts. MULTI system is better than the AU, UK and US models but inferior to the UNIV system. We observed that the multiple pronunciation for individual words in the lexicon may not be the right choice in accelerating the system robustness due to accent variations. Overall UNIV outperforms all the other models and yields about 73%, 65%, 62% and 8% in string error rate reductions when compared to the US, AU, UK and MULTI systems

Data Model	LL_US	
	String Accuracy	Arc Count
AU	53.6%	30079
UK	64.5%	31094
US	93.6%	19675
MULTI	91.5%	88257
UNIV	93.1%	23818

Table 4: String accuracy and arc-count for a known-length connected-digit recognition task using landline American English (LL_US) data as a function of various model type.

Model Type	Model Size	String Accuracy	Average Arc Count
AU	1.54MB	75.0%	21343
UK	1.50MB	76.8%	22551
US	1.36MB	67.4%	19869
MULTI	4.40MB	90.4%	75221
UNIV	2.60MB	91.2%	20931

Table 5: Model size, average string accuracy and arc-counts across LL_AU, LL_UK and LL_US databases as a function of model type.

respectively. UNIV exhibits consistent improvements on every LL_AU, LL_UK and LL_US databases. Furthermore, the string accuracies are in the low 90% in all three databases and this suggests that the universal acoustic models can be used for real telephony applications. The average arc count is three times less than the MULTI system and comparable with the best monolingual systems. Also the model size is twice that of the AU, UK and US models but half of that of MULTI system. From the table, it is clear that the UNIV model significantly outperforms the other systems in most cases, as expected. To conclude, the UNIV system provides an efficient way of modeling accent variations from the three languages by using a single Viterbi decoder. It is encouraging that our goal of designing a single global system for all three languages (Australian, American and British English) is achieved by using the universal hybrid system, and the test results have demonstrated the efficacy of enhanced hybrid system.

5. Conclusions

This paper addressed the problem of speech recognition through regional accents. Universal hybrid modeling system has been proposed and investigated in this paper by intelligently combining data from many different accented speech. The experimental results showed that the universal model in conjunction with suitable model topology to represent the extraneous speech accents not only provide good recognition accuracy but also yield faster response with reduced model size. The major benefit of using an universal hybrid system for a particular language is that there is no need to know about the prior knowledge concerning the nature of the speaker accent that exist in the modern telephone network. In the future experiments, we will apply this universal technique for other languages such as French (Canadian and Continental) and Spanish (Castilian, Mexican and Columbian) in the system.

Acknowledgements

The author would like to thank Dr. Rafid Sukkar for helpful discussions and support in the early stages of this work.

6. References

[1] L. M. Arslan and J. H. L. Hansen, “Frequency Characteristics of Foreign accented speech”, *Proc. ICASSP*, pp. 1123-1126, 1997.

[2] R. Chengalvarayan, “A comparative study of hybrid modelling techniques for improved telephone speech recognition”, *Proc. ICSLP*, pp. 313-316, 1998.

[3] R. Chengalvarayan, “Use of multiple classifiers for speech recognition in wireless CDMA network environments”, *Proc. ICSLP*, Vol. 3, pp. 386-389, 2000.

[4] R. Chengalvarayan, “Hybrid HMM architectures for robust speech recognition and language identification,” *Proc. Systemics, Cybernetics and Informatics*, Vol. 6, pp. 5-8, 2000.

[5] V. Fisher, Y. Gao and E. Janke, “Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer”, *Proc. ICSLP*, 1998.

[6] J. L. Hieronymus and S. Kadambe, “Robust spoken language identification using large vocabulary speech recognition”, *Proc. ICASSP*, pp. 1111-1114, 1997.

[7] C. H. Ho, S. Vaseghi and A. Chen, “Voice conversion between UK and US accented English”, *Proc. EUROSPEECH*, Vol. 5, pp. 2079-2082, 1999.

[8] S. Katagiri, B-H. Juang and C-H. Lee, “Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method”, *Proc. IEEE*, Vol. 86, No. 11, pp. 2345-2373, 1998.

[9] J. B. Puel and R. Andre-O’brecht, “Cellular phone speech recognition: Noise compensation versus robust architectures”, *Proc. EUROSPEECH*, pp. 1151-1154, 1997.

[10] J. Sturm and E. Sanders, “Modelling phonetic context using head-body-tail models for connected-digit recognition”, *Proc. ICSLP*, pp. 429-432, 2000.

[11] R. A. Sukkar, A. R. Setlur, C-H. Lee and J. Jacob, “Verifying and correcting recognition string hypotheses using discriminative utterance verification”, *Speech Communication*, Vol. 22, pp. 333-342, 1997.

[12] A. C. Surendran, C-H. Lee and M. Rahim, “Nonlinear compensation for stochastic matching”, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 6, 1999, pp. 643-655.

[13] D. L. Thomson and R. Chengalvarayan, “Use of periodicity and jitter as speech recognition features”, *Proc. ICASSP*, pp. 21-24, 1998.

[14] F. Weng, H. Bratt, L. Neumeyer and A. Stolcke, “A study of multilingual speech recognition”, *Proc. EUROSPEECH*, pp. 359-362, 1999.

[15] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech”, *IEEE Transactions on Speech and Audio Processing*, Vol. 4, No. 1, pp. 31-44, 1996.

[16] The European Language Resourece Association web site: *http://www.icpgrenet.fr/ELRA*