Interspeech 2018
2-6 September 2018, Hyderabad, India

Hierarchical Accent Determination and Application in a Large Scale ASR System

Ramya Viswanathan, Periyasamy Paramasivam, Jithendra Vepa

Samsung R&D Institute, India - Bangalore

{r.vishwanath, periyasamy.p, jithendra.v}@samsung.com

Abstract

In deploying Automatic Speech Recognition Systems (ASR) on a global scale, several challenges arise for supporting a widely used language such as English. The primary one among them is to deal with a wide variety of accents. We propose a Hierarchical Accent Determination system that deals with accent variations across large geographical regions at macro level and then the variations at the sub-regions within a selected large geographical region at micro level along with taking context cues. Eight accents [GB, US, Australian, Canadian, Spanish, Korean, Indian & Chinese] are identified at macro level and accent-specific models corresponding to the identified accents are used. The accuracy of the accent identification system is around 80% with ASR as well as using context cues such as phone language and keyboard language. The deployment of the accent identification system has improved the overall accuracy of Speech Recognition system by 10% for accented speech. It is planned to expand the approach to identify accents with significant variations found at sub-regional level in India such as Hindi, Tamil, Telugu, Malayalam, and Bengali.

Index Terms: speech recognition, accent determination, contextual speech recognition

1. Introduction

Several approaches have been proposed in the literature for accent compensated Speech Recognition Systems. First approach is to have an accent independent Acoustic model as proposed by Chengalvarayan[1]. However, the problem with this approach is that speech data for all the different accents are required prior hand for training and it is not flexible enough to include context information such as location, keyboard language, contact name languages etc., which are readily available when deployed in systems like mobile phones. The second approach is to run several accept-specific recognizers in parallel and selecting the one with highest likelihood as proposed in [2]. However as commercial deployments start to move more towards on-device recognition for commercial reasons, running multiple recognizers on devices such as mobile phones is not practically possible due to memory, computation speed and power constraints. The third approach is to let an independent Accent Identification system to determine the accent as proposed in [2] and using it to choose the correct accent-dependent recognizer once accent is identified. We have chosen the independent Accent identification approach as our base and it does not have some of the key problems identified above in the other approaches. First, since our target is to apply for mobile devices which are basically single user devices, we do not need to identify the accent every time user speaks to the device. Identifying the accent once reliably and choosing the most suited accent model is sufficient which is what exactly done by our approach while the first two approaches identify accent every time wasting computational resources. Second, from the deployment point of view, our method is flexible as the system can be extended by adding new accents without disturbing the existing system while Accent-independent model needs to be rebuilt every time a new accent has to be included or accuracy of an accent needs to be improved. Third as mentioned before, a lot of context cues are available in mobile devices which cannot be incorporated easily in the first two approaches as it has been done in our approach. In the next sections, we will describe our system architecture, method to choose words with maximal accent distance and test results.

2. System Architecture

Our overall system architecture is shown in figure 1. A set of five sentences is selected as described in the section 2.1 and user is asked to read them. All the five sentences are first sent to the first level of accent specific recognizers, in our case, eight of them – Korean, Spanish, Indian, Korean, US, Australian, Great Britain and Canadian. The recognizers are complete speech processing systems which produce recognized text as output. WER rate is calculated for each accent from the output text. A weight of w1 is added for the model corresponding to the country in which the mobile is used, a weight of w2 is added for the model corresponding to the phone language and a weight of w3 added for the model corresponding to the keyboard language. In case if any of the phone, keyboard or country language falls outside of these eight models, then more weight is added to US model as default. The model which comes out with top score after all the weight additions done is taken as first level accent. In case if there is no further sub-accent classification as in the case of Canadian or Australian accent, the accent is sent to server and server sets the accent accordingly.

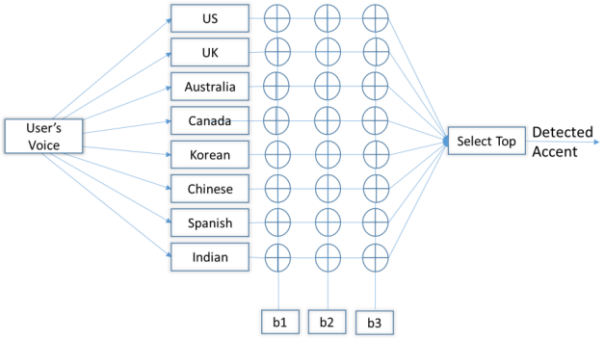

Figure 1: *System Architecture.*

However, in cases where further sub-accent classification is possible, another five sentences to differentiate between the sub-accents are shown to the user which he/she needs to read. WER rate is calculated for each sub-accent. A weight of w4 and w5 are added to the models corresponding to phone language and keyboard language respectively. The accent with the top score is sent to the server as selected accent. The server selects the accent specific model and associates with the device. From then onwards, all the queries from the device are handled using the same selected model at the server.

The weights w1, w2, w3, w4, w5 are calculated relative to the top model accuracy according to the formula

$$w_n = (100 - Top) * 0.25 + c_n$$

Where c_n are constants that define relative weightages to country language, phone language and keyboard language. The relative weightages for country language, phone language and keyboard language are given values in the decreasing order based on empirical study.

2.1. Sentence selection

The sentences are selected with specific words which will have distinct pronunciation for a particular accent group as shown in Table 1. For example, pronunciation of "very" and "Festival" by a Korean native speaker will be different from others due to mapping of '/v/' sound to '/b/' and '/f/' sound to '/p/' [3]. "Zoom" is pronounced differently by a person with Spanish accent as '/z/' sound does not exist in Spanish [4]. Similarly, Mandarin speakers tend to delete '/l/' sound in multisyllabic words such as "Usually" [5]. In the same way, "Genre" is pronounced differently by a person with Indian accent. Using this linguistic knowledge, the sentences are chosen such that there will be a considerable difference among all accents that will reflect as error with accent-specific models other than the expected one.

Table 1: *Sentences with Dominant Accents*

Key Word	Accent	Utterance
Usually	Chinese	It **usually** snows in the winter
Genre	Indian	play some music from rock **genre**
Very, festival	Korean	A very nice festival
Zoom	Spanish	**Zoom** in on this picture

In the similar manner, another set of five sentences are chosen for every sub-accent identification task and presented to user to read.

3. Experimental Results

The performance of the system has been evaluated from custom dataset with data collected from different regions with atleast 100 speakers for each accent. Table 2 shows the accent determination accuracy with and without context cues. As we can see from the table, the average accuracy of accent determination system comes to 50% without any context cues. However, it goes up to 80% while considering the context cues such as native language of the user's region, user's phone language and keyboard language.

Table 3 shows the distribution of accents identified across countries collected over a certain period of usage. It can be seen from the table that the distribution of accents is roughly proportional to the relative distribution of the ethnicity in the population. However, one may notice US accent being disproportionately used from the table. But it is due to the fact that the system is tuned to US accent by default if it is not very sure about the non-US accent.

Table 2: *Accent Determination Accuracy*

Language	ASR Accent Determination	
	w/o Context Cues	With Context Cues
Indian	46.3	93.7
Chinese	65.6	93.8
Korean	41.0	69.9
Spanish	51.9	65.4
US	94.5	95.4

From the word accuracy of accent specific model and the distribution of usage of accent specific models in different countries, it is estimated that overall accuracy of the system has improved by over 10%.

Table 3: *Distribution of Accent Models Usage*

Country	Accent Models							
	en-US	en-AU	en-CA	en-CN	en-ES	en-GB	en-IN	en-KR
USA	86.98%	0.01%	0.01%	1.20%	3.76%	0.01%	5.05%	2.99%
IND	11.89%	0.08%	0.04%	0.12%	0.38%	0.05%	86.85%	0.59%
CAN	72.20%	0.02%	12.33%	1.22%	2.37%	4.19%	5.72%	1.95%
GBR	47.19%	0.02%	0%	1.19%	1.52%	45.50%	2.75%	1.83%
AUS	35.19%	53.92%	0.01%	1.05%	1.08%	2.77%	3.61%	2.38%

4. Conclusions

We have demonstrated an accent determination system that has been integrated with a large scale speech recognition task. A specific approach of independent accent identification system has been chosen considering the demands of deploying a large scale system. The performance of each model is tested for different accents and distribution of model use in a large population is captured and presented. The deployment of the system has improved the overall ASR accuracy by 10% compared to deploying a single model for all the accents. Further work is in progress to identify sub-accents among Indian accents such as Hindi English, Tamil English, Telugu English, Malayalam English etc.

5. References

[1] R.Chengalvarayan, "Accent-independent universal HMM-based speech recognizer for American, Australian and British English" *Proceedings of Eurospeech*, 2001, pp. 2733-2736

[2] A.Faria, "Accent classification for speech recognition" *Proceedings of the Second International Workshop on Machine Learning for Multimodal Interaction,* 2005, Edinburgh, UK: Springer, 2005.p.285-293.

[3] J.Kim, "Second Language Acquisition:Phonology" *The Handbook of Korean Linguistics*, L.Brown and J.Yeon, Wiley,2015.p.373-387

[4] M.McDonald, "The influence of Spanish phonology on the English spoken by United States Hispanics" *American Spanish pronunciation: Theoretical and applied perspectives,* 2004, Georgetown University Press, pp. 215–236.

[5] D.Deterding, "The pronunciation of English by speakers from China" *English World-Wide*, vol.27, no.2, pp.175-198, 2006