http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 5 (6): 1140-1142, 2006
ISSN 1812-5638

Design of Language Independent Speech Interfaces

[1]Yasir Khalil Ibrahim and [2]Mazin M. Al Hadidi
[1]Department of Computer, Jerash Private University, Jordan
[2]Department of Computer Engineering, Al Balqa University, Jordan

Abstract: Speech recognition technology has made spoken interaction with machines feasible. However, no universal interface has yet been proposed for human to communicate effectively, efficiently and effortlessly with machines via speech. Since the existing speech interfaces are language specific and language dependent, it is proposed to standardize the interface to make it language independent and adaptive to local dialects. The two existing primary models of speech recognition (Sehultz and Waibel, 2001) the acoustic model (analyses the sound of voice and convert them to phonemes) and the language models (compares the combination of phonemes to the words in its digital dictionary)-are language and speaker dependent. The extension of these models, Large Vocabulary Continuous Speech Recognition, is proposed in this study. Monolingual recognizers for multiple languages are designed first and the entire collection of phoneme set is called as GlobalPhone database. Based on this global unit set, it is aimed to make the resulting multilingual acoustic model as language independent. The phonemes with common sound are refined from GlobalPhone database to form a set of International Phonetic Alphabets (IPA). From this, the phonemes can be derived for any target language. To make the interface adaptive to different dialects, phoneme models of arbitrary context width called polyphones from resulting target language are maintained. By evaluating the Polyphone Decision Tree Specification, which is context-relative of that target language, the recognizer can be adapted to all accents and dialects of that particular language.

Key words: Speech recognition, pattern recognition, language processing, decision trees

INTRODUCTION

Speech in general is not a direct replacement to keyboard or barcode, today's primary data entry devices. However, speech is generally better than typing when the work requires the operator be hands-free and/or eyes-free. The latest speech recognition technologies work well in noise and accents do not affect the accuracy, since each user trains the system. But to develop such state of art products, a lot of background works need to be done. The data dictionary for that language, the pronunciation rules are resources that need to be defined for the model to be accurate. Such systems require lengthy development, which is data and labor intensive and heavy involvement by experts who meticulously craft the vocabulary, grammar and semantics for specific domain. So, no universal interface has yet been proposed for human to communicate effectively, efficiently and effortlessly with machines via speech (Sehultz and Waibel, 2000). The aim of this study is to standardize the current working models so that they can be extended to any target language and its various dialects. A multilingual recognizer model that can be derived collectively from the existing models for popular languages is proposed in this study.

Existing speech recognition models: In the existing model (Douglas, 2001), the speech signal is received from an input device (example: A microphone) and converted from analogue to digital information. After digitization, the signal is classified into a set of codes that the system can comprehend. This input data can be processed using two primary component of speech recognition. The first component, called the acoustic mode, analyzes the sound of your voice and converts them to phonemes, the basic elements of speech. The English language contains approximately 50 phonemes. The acoustic model removes noise and unneeded information such as changes in volume in the first phase. Then, using mathematical calculation, it reduces the data to a spectrum of frequencies (the pitches of sounds), analyses the data and converts the words into digital representation of phonemes.

The second component, called the language model, analyses the contents of speech and compares the combinations of phonemes to words in its digital dictionary that is a huge database of the most common words in the English language. The language model then search through the dictionary to extract the spelled words. This method of speech recognition that works with large

Corresponding Author: Dr. Yasir K. Ibrahim, Department of Computer, Jerash Private University, Jordan

databases and phonemes analyzers is called Large Vocabulary Continuous Speech recognition (LVCSR).

To build a typical recognizer (Jurafsky and Martin, 2000), this data is usually includes a large volume of recorded and transcribed speech. Unfortunately, the assumption that large speech databases can be provided on demand does not hold for several reasons.

- The collection of large databases requires a tremendous amount of time and resources.
- More than 4500 languages exist in the world and only 150 languages have more than one million speakers. The reset of languages have less than 100,000 native speakers. So, to design recognizers for these languages using the existing models becomes a Herculean task.

The GlobalPhone database: GlobalPhone is a high-quality read speech and text database in a large variety of languages that is suitable for the development of Large Vocabulary Speech recognition Systems in many languages. The need for developing such a database has a number of reasons. As the demand for rapid deployment of LVCSR in many languages grows, new approaches for cross-language transfer like the development of language independent GlobalPhone sets and language adaptive speech recognizers are of increasing concern. These interest accompanied by the need for a multilingual speech and text database that covers many languages and is uniform across languages. Uniformity here refers to the total amount of text and audio per language as well as to the quality of data, such as recording conditions (noise, channel, microphone, etc.), collection scenario (task, setup, speaking style, etc.) and the transcription conventions. Only uniform data allow the development of GlobalPhone sets and enable the comparison of speech and/or text across languages. Furthermore, the research on language independent and language adaptive speech recognition requires databases that cover the most relevant languages.

The languages with the largest number of speakers arranged in rank order as follows: English, Hindi, Spanish, Russian, Arabic, Bengalese, Portuguese, Malay, Japanese, French, German and Korean as given in Webster's New Encyclopaedia Dictionary.

Out of these languages, the languages given in italics were chosen for the GlobalPhone database. The data acquisition session consisted of recorded data from about 100 native speakers each reading 15 min of text. The speakers were chosen from both sexes, adults of various ages and educational levels. As a result of the uniform collection procedure, it is suitable for language independent and language adaptive LVCSR as well as for language identification tasks.

Multilingual acoustic modelling: As the demand for speech recognition systems in multiple languages grows, the development of multilingual systems which combine the phonetic inventory of many languages (GlobalPhone database) into one single acoustic model set is of increasing importance. For multilingual speech recognition, it is intended to share acoustic models of similar sounds across the languages. Similarities of sounds are given in the Appendix using WorldBet notation.

Altogether there are 78 phonemes plus a silence and two noise models for spontaneous speech effects. 14 phonemes are shared across all five languages, but half of the set consists of monophonemes belonging to only one of the five languages. Based on these 78 phonemes categories, three different multilingual systems, namely ML5-mix, ML5-sep and ML5-tag, are built. In ML5-mix, for each of the 78 phonemes one mixture of 16 Gaussian distributions is initialized and train the models by sharing the data of all five languages. The resulting recognizer ML5-mix is a fully continuous system with 3000 models mixed over all languages. In ML5-sep each element is modelled separately for each language. No data are shared; all models except silence and noise are a language dependent. For each of the 170 phonemes, one mixture of 16 Gaussian distribution is initialized, after training this results ML5-sep is fully continuous system with 3000 language dependent models. In ML5-tag, a language tag is attached to each of the 78 phoneme categories in order to preserve language information.

International Phonetic Alphabets (IPA): Sounds production is human, not language specific is the basic concept of voice recognition. The data collected into the

Appendix: Similarities of sound using world bet notation

Phonemes (World Bet)	K	S	C	T	J	Sum
N,m,l,ts,p,b,t,d,g,k	Y	Y	Y	Y	Y	
I,e,o	Y	Y	Y	Y	Y	14
F,j,z		Y	Y	Y	Y	
R,u	Y	Y	Y	Y		
Dz	Y		Y	Y	Y	6
A	Y	Y	Y			
S			Y	Y	Y	
H	Y			Y	Y	
4	Y	Y			Y	4
W,x,L		Y	Y			
A				Y	Y	
N	Y	Y				
V,z			Y	Y		
Y,7	Y			Y		
ts			Y		Y	10
P',t',k',dz',s',oE,oa,4i,ue,E,O,I,U,iu,ie,io,ia	y					17
D,G,T,V,r9,ai,au,ei,eu,oi,a+,e+,I+,o+,u+		y				15
Palatal c, palatal d			Y			2
Ix,soft X2				Y		2
?,Nq,V:a:,e:,i:,o:,4:					Y	8
Monolingual Sum = 170	40	40	30	29	31	
Multingual						78

K-Korean, S-Spanish, T-Turkish, J-Japanese, C-Croatian

GlobalPhone database from different languages leads to this basic concept. So, it is proposed to extract the phonemes with common Sound from the database and group them as International Phonetic Alphabets (IPA).

Also it is proposed to maintain a universal sound inventory based on the IPA sounds, which count to 162-IPA sound classes. Each sound class is represented by one phoneme that is trained through data sharing across languages.

- m, n, s, l occur languages.
- P, b, t, d, k, g, f and I, u, e, a, o occur in almost all languages.

Thus using the IPA set, the phones can be derived for any target languages.

Polyphone decision tree specification: Assume that the acoustic model for a target language has been derived, but however to make the acoustic model complete, the model has to be modified so as to suit the context of the particular language.

To achieve context dependant phoneme models, a decision tree clustering procedure is applied. This procedure uses an entropy based distance measure defined over the mixture weights of Gaussians and a question set that consists of linguistically motivated question about the phonetic context of a phoneme model. During clustering, the question with the highest entropy gain is selected when splitting same IPA symbol. After reaching the predefined number of polyphones the splitting procedure is terminated. This clustering procedure is extended to the multilingual case by introducing question about the language and language groups to which a phoneme belongs. That is, the decision whether phonetic context information is more important than language information becomes data driven. The models were tested with large volume of quintphones over the five different languages as continuous system.

CONCLUSIONS

In this study, the design of adaptive acoustic speech interface for any target language is discussed. This requires a considerable amount of data from the target language and works with the resources from the Globalphone database. The resulting target language recognizer can be made as adaptive using the concept of polyphone Decision Trees.

ACKNOWLEDGMENTS

This work has been adopted at the University of Jerash, Jordan (Programming part) and at the University of Balqa (Application part).

REFERENCES

Douglas, O.S, 2001. Speech Communication-Human and Machine. 2nd Edn., University Press.

Jurafsky, D. and J.H. Martin, 2000. Speech and Language Processing. Pearson Education.

Schultz, T. and A. Waibel, 2000. Polyphone decision tree specialization for language adaption. IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey.

Schultz, T. and A. Waibel, 2001. Language independent and language adaptive acoustic modeling. Speech Communication, 35: 31-51.