IT in Industry, vol. 3, no. 1, 2015 Published online 31-Mar-2015

Speech Feature Extraction and Data Visualisation

Vowel recognition and phonology analysis of four Asian ESL accents

David Tien
School of Computing and Mathematics
Charles Stuart University
Bathurst, NSW, Australia

Yung. C. Liang and Ayushi Sisodiya
Department of Electrical and Computer Engineering
National University of Singapore
Singapore

***Abstract*—This paper presents a signal processing approach to analyse and identify accent discriminative features of four groups of English as a second language (ESL) speakers, including Chinese, Indian, Japanese, and Korean. The features used for speech recognition include pitch, stress, formant frequencies, the Mel frequency coefficient, log frequency coefficient, and the intensity and duration of vowels spoken. This paper presents our study using the Matlab Speech Analysis Toolbox, and highlights how data processing can be automated and results visualised. The proposed algorithm achieved an average success rate of 57.3% in identifying vowels spoken in a speech by the four non-native English speaker groups.**

Keywords—speech recognition; feature extraction; data visualisation;vowel recognition; phonology analysis

I. Introduction

In speech processing, pattern recognition is a very important research area as it helps in recognizing similarities among different speakers, and plays an indispensable role in the design and development of recognition models and automatic speech recognition (ASR) systems. Pattern recognition consists of two major areas: feature extraction and classification. All pattern recognition systems require a front end signal processing system, which converts speech waveforms to certain parametric representations, called features. The extracted features are then analysed and classified accordingly [1, 2]. Feature extraction techniques can be broadly divided into temporal analysis and spectral analysis. In temporal analysis, the speech waveform is used directly for analysis; while in spectral analysis, spectral representation of the speech signal is used instead. In our study, temporal analysis techniques employed include pitch, intensity, formant frequency, and log energy; whereas spectral analysis techniques employed include the Mel-frequency spectral and log-frequency spectral.

II. Vowel Recognition and Phonology Analysis

This paper proposes an easy to understand and user friendly approach for speech feature extraction and data visualization. Software in Matlab has been developed to recognize features as spoken by four groups of non-native English speakers (or English as Second Language, ESL) – Chinese, Indian, Korean, and Japanese. These groups of speakers have very distinctive accents.

A. Matlab – Speech Analysis Toolbox

The developed software makes the speech processing tasks easy and accurate. The language toolbox provides the user with a general envelop graph for different languages, and is very useful for the development of speech recognition applications. This work also plays a vital role as a catalyst for future linguistic research.

The major functionalities of the software used in our work include:

- Pitch Analysis
- Intensity Analysis
- Frequency Analysis
- Log Energy Analysis
- Log- Frequency Power Analysis
- Mel- Frequency Cepstral Analysis
- Vowel Identification
- Recording and maintaining of Speech Corpus of different languages
- Language analysis of the speech features of different languages

B. Formant Frequency for Vowel Identification

The formant frequency is considered a very important aspect of the speech signal in the frequency domain. We used formant frequencies in our work for vowel identification. We will discuss the algorithm and its performance later in this paper.

As explained by Prica and Ilić, the main difference between vowels and consonants is that vowels resonate in the throat. When a vowel is pronounced, the formants are exactly the resonant frequencies of the vocal tract [3]. It is a natural and obvious choice to use formant frequencies for vowel identification.

The frequencies and levels of the first three formants of vowels were measured. The statistical analysis of these formant variables confirmed that the first three frequencies are the most appropriate distinctive parameters for describing the spectral differences among the vowel sounds. Maximum

  ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

likelihood regions were computed and used to classify the vowels. For speech recognition in particular, the formant analysis is more practical because it is much simpler and can be carried out in real time. The steps in our implementation are listed below:

1. Read a speech file in wav format and store it into a vector.
2. Measure the sample frequency and calculate the pace of the speech according to which it divided the speech into appropriate fragments for formant analysis.
3. Select the small divided signals one by one, and further divide them into 10 equal blocks in the time domain.
4. Select the block with maximum power content, which is also a measure of the stress of vowel.
5. Normalize the selected block and shows the waveform of the vowel segment.
6. Determine the frequencies at which the peaks in the power spectral distribution occur.
7. Extract and compare the first three formants with the segment extracted.
8. Calculate the Euclidean distance between the set of frequencies obtained from the user, and each of the set of frequencies corresponding to the vowels.
9. Use the minimum distance criterion to determine the vowel.

The vowel pronunciations taken are listed below with their Formant frequency which are calculated through the formant frequency code:

- I ≈ IY = [255 2330 3000];
- I ≈ IH = [350 1975 2560];
- E ≈ EH = [560 1875 2550];
- A ≈ AE = [735 1625 2465];
- A ≈ AA = [760 1065 2550];
- O ≈ AO = [610 865 2540];
- U ≈ UW = [290 940 2180];
- U ≈ UH = [475 1070 2410];
- A ≈ AH = [640 1250 2610];

TABLE I. VOWELS POTENTIALLY SPOKEN BY NON-NATIVE ENGLISH SPEAKER GROUPS

it's	a	pe	ace	ful	ex	is	te	en	ce
IY	UH	EH	AE	UH	EH	IY	EH	EH	IH
IH	AH	IH	IH	UW	AE	IH	UH	UH	IY
EH	AA	~	EH	~	~	EH	~	IH	~
~	~	~	~	~	~	~	~	IY	~

C. Vowel Identification Results

We compared the similarities as well as differences in the way vowels are spoken by each of the four groups of speakers. We present the performance results of the proposed vowel identification algorithm and findings of our phonology analysis below. The sample speech used in the tests was,

"It's a Peaceful Existence."

In this sentence the possible vowel sounds to be recognized are shown in Table 1. As non-native English speakers have their own accents, we expected to recognise these vowels due to potential inaccurate or incorrect pronunciations of the speakers.

The algorithm mentioned above was run on all ten speech samples for each of the four speaker groups, totalling forty speech samples in the corpus. The success rate was calculated using the matching percentage of results. One example for each of the groups, after running the code, for the first vowel "It's" (IY, IH, EH) is given in Table 2 to Table 5. The success rates for recognising 'It's' as vowels are 83.3%, 86.6%, 33.3%, and 46.1% for Chinese, Indian, Japan, and Korean speaker groups, respectively.

Similarly, the success rates of recognising all the vowels in the sentence for all forty speech samples for each speaker group are summarised in Table 6.

D. Observations of Individual Languages

From the above results, our algorithm is the most successful in recognising the vowels 'a' and f'u'l. The main reason was that these syllables are spoken with the highest stress in the sentence. Our algorithm is the least successful in identifying the vowel sounds 'ex' and 'pe'. The main reason was that the Euclidean distances of these syllables were more similar to the vowels immediately after ('is') or before ('a') them, which misled the algorithm to choose the wrong vowels instead. For the 40 speech samples the average success rates is ≈ 57.3%. As the number of speech samples increases, the probability of successful identification by the algorithm also increases, hence increasing the success rate.

TABLE II. RESULTS OF CHINESE SPEAKERS

Speaker	Vowel Detected	Matching? (yes/no)	Matching Percentage
Chinese 1	IY	Yes	100%
Chinese 2	EH	Yes	100%
Chinese 3	IY, IY	Yes	100%
Chinese 4	EH, EH	Yes	100%
Chinese 5	IY	Yes	100%
Chinese 6	No Vowel	No	0%
Chinese 7	IY	Yes	100%
Chinese 8	IH	Yes	0%
Chinese 9	UH	No	100%
Chinese 10	EH	Yes	100%

 ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

TABLE III. RESULTS OF INDIAN SPEAKERS

Speaker	Vowel Detected	Matching? (yes/no)	Matching Percentage
Indian 1	IY	Yes	100%
Indian 2	IY	Yes	100%
Indian 3	IY,IH	Yes	100%
Indian 4	IY	Yes	100%
Indian 5	IY,IY,IY	yes	100%
Indian 6	IY,IY	yes	100%
Indian 7	IY,	yes	100%
Indian 8	IY	yes	100%
Indian 9	AE, IH	Partial yes	50%
Indian 10	No vowel, IH	Partial yes	50%

TABLE IV. RESULTS OF JAPANESE SPEAKERS

Speaker	Vowel Detected	Matching? (yes/no)	Matching Percentage
Japanese 1	IY, No Vowel	Partial Yes	50%
Japanese 2	UH, No Vowel	No	0%
Japanese 3	UH, No vowel	No	0%
Japanese 4	EH, IY	Yes	100%
Japanese 5	AH, UH	No	0%
Japanese 6	No vowel, UH	No	0%
Japanese 7	IY	Yes	100%
Japanese 8	UH, UH	No	0%
Japanese 9	IY	Yes	100%
Japanese 10	IY, UH	Partial yes	50%

TABLE V. RESULTS OF KOREAN SPEAKERS

Speaker	Vowel Detected	Matching? (yes/no)	Matching Percentage
Korean 1	IY	yes	100%
Korean 2	No Vowel	No	0%
Korean 3	EH,	Yes	100%
Korean 4	IY	Yes	100%
Korean 5	UH	No	0%
Korean 6	IY,UH	Partial Yes	50%
Korean 7	UH	No	0%
Korean 8	No vowel, UH	No	0%
Korean 9	IY, IY	Yes	100%
Korean 10	UH	No	0%

TABLE VI. SUCCESS RATES OF VOWEL IDENTIFICATION ALGORITHM

Speaker	its	a	pe	Ace	Ful	Ex	is	te	en	ce
Korean	46.1%	100%	10%	30%	75%	55%	43%	70%	70%	64%
Chinese	83.3%	86.2%	20%	64%	87%	34%	20%	60%	60%	20%
Japanese	33.3%	60%	45.5%	56%	73%	0%	68%	60%	60%	68%
Indian	86.6%	80%	40%	60%	68%	46%	55%	30%	30%	56%
Average Success Rate	62.6%	81.55%	42.95%	52.5%	75.75%	38.75%	51.5%	57.5%	57.5%	52%

TABLE VII. STRESS ENENGY LEVEL OF A CHINESE SPEAKER

Vowel Detected	Stress Energy(db)	Time of occurrence
EH	0.000454969	0.1 sec
EH	0.000611704	0.281429 sec
UW	0.000880056	0.46 sec
IY	0.000850723	0.644286 sec
IY	0.000387354	0.725714 sec
AH	0.000483724	1.00714 sec
IY	0.00086256	1.18857 sec
IH	0.000555674	1.37 sec
IY	9.24567e-005	1.55 sec
IY	0.000111215	1.732 sec
No vowel		
UW	0.000183622	2.09571 sec
UW	0.00019356	2.27714 sec
UW	0.000205966	2.45857 sec
IY	0.000118864	2.64 sec
IY	0.000835543	2.72143 sec
IH	0.000378061	3.00286 sec
IH	7.88983e-005	3.18429 sec
UW	0.000179787	3.36571 sec

After looking at the overall results, we have further analysed the following three aspects of each of the four speaker groups:

1. Vowels identified, with stress levels and the times of occurrence. An example of a Chinese speaker is shown in Table 7.
2. Normalised signal graphs of the individual vowels. The graphs of all four speaker groups are shown in Fig. 1, 3, 5, and 7.
3. Stress energy graphs of the individual vowels. The graphs of all four speaker groups are shown in Fig. 2, 4, 6, 8.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

1) Chinese

- For Chinese speakers, stressed syllables tend not to stand out, due to the fact that fully-stressed syllables often occur together in the Chinese language.
- They pronounce the /UW/ /UH/ /EH/ for a long duration (Vowel a).
- There is no distinction between /EH/ and /IY/

2) Indian

- The general pattern of the Indian speakers is that their vowels are spoken clearly, and stress on /UW/ is high compared to the other languages.
- The distinction between /IY/ and /UH/ is minimal.
- More vowel sounds were detected, indicating the duration for spoken vowels is longer.

3) Japanese

- Japanese speakers tend to spend equal time on each syllable when speaking English, due to the fact that syllables occur at regular time intervals in the Japanese language.

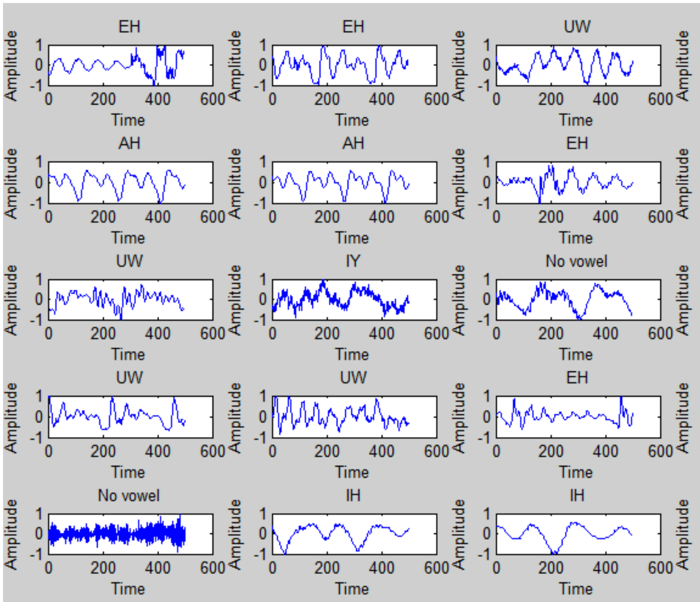

Fig. 1. Normalized Signals of the Vowels Identified of Chinese Speaker

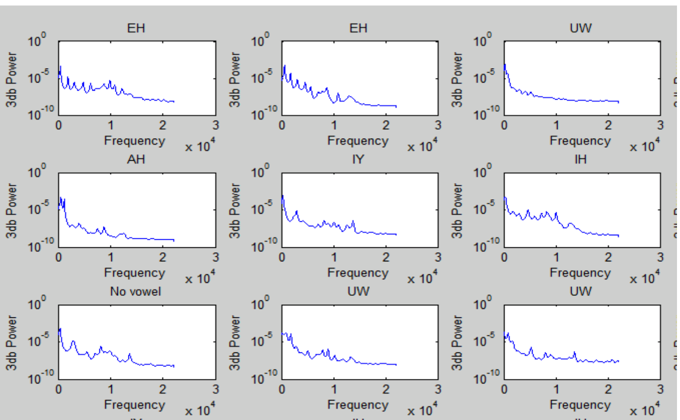

Fig. 2. Stress Intensity Graph of the Vowels Identified of Chinese Speaker

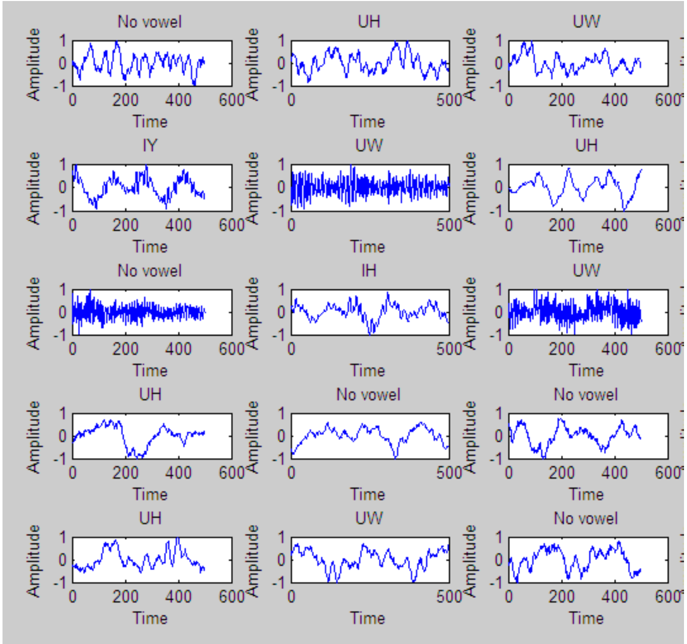

Fig. 3. Normalized Signals of the VowelS Identified of Indian Speaker

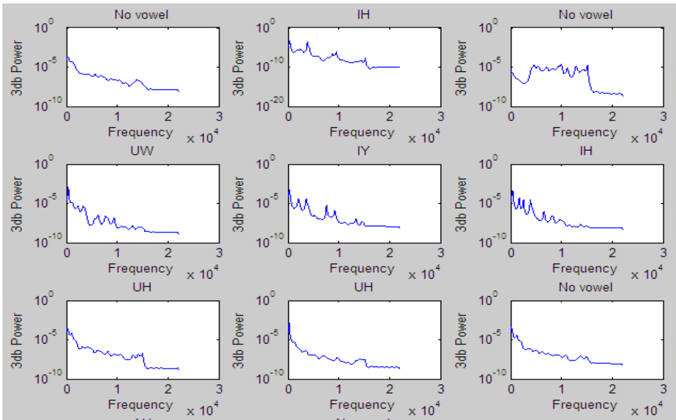

Fig. 4. Stress Intensity Graph of the Vowels Identified of Indian Speaker

- Vowel A is pronounced as /EH/ and never like /UW/
- Similar to other speakers there is no distinction between /EH/ and /IY/

4) Korean

- Koreans in general speak every character very fast, thus resulting in sounding of vowels very similar.
- Stress level is high on /IY/
- There is no distinction between /UH/ and /IY/
- There is no distinction between /EH/ and /IY/

E. Phonlogy Analysis

Observations made on the four languages are presented below. Our findings are generally consistent with existing knowledge in linguistic research.

1) Chinese Speakers

- Chinese speakers find it difficult to differentiate between the /l/ and /r/

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

- Chinese speakers often cannot pronounce /t/ ,instead they say it as a /θ/
- Chinese speakers often cannot differentiate between /s/ and /ʃ/
- In English, stressed syllables often stand out. In Chinese, full syllables often occur together, so stressed syllables often do not stand out [4]. When speaking in English, Chinese speakers often do not put enough emphasis on stressed syllables.
- Chinese speakers tend to have unnecessary aspiration of plosives after /s/.
- Chinese speakers tend to speak the long paired vowels and short pair vowels in a similar way, for example, /i/ in did and /i:/ deed are normally pronounced as the same sound-long vowel /i/.
- Both /u/ and /u:/ in pull and pool are normally pronounced as long vowels sound /u/.
- Chinese speakers tend to replace/h/ with /x/ in most cases even though they sound very different.

2) Indian Speakers

- Indian speakers use /e/, /o/ complementarily with the more common /i/, /u/.
- Indian speakers tend to pronounce /a/ in variation with the rounded and more back /ɒ/.
- Indian speakers cannot make a distinction between /ɒ/ and /ɔ:/.
- Indian speakers have a more American style of talking. They pronounce /a/ instead of rounded / ɒ/ or /ɔ:/.

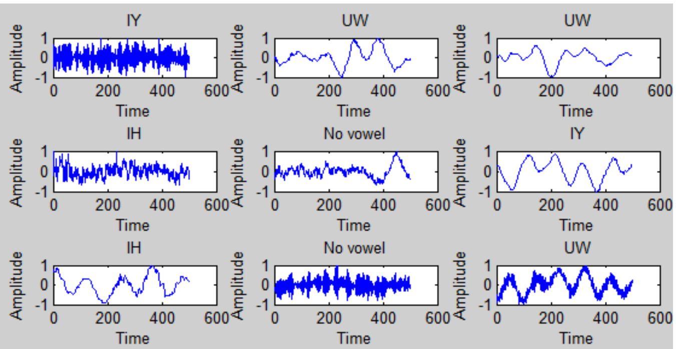

Fig. 5. Normalized Signals of the Vowels Identified of Japanese Speaker

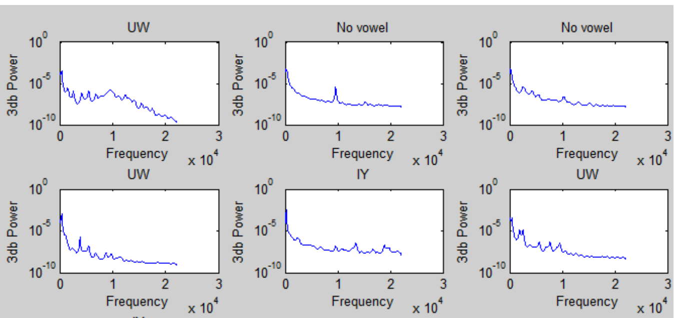

Fig. 6. Stress Intensity Graph of the Vowels Identified of Japanese Speaker

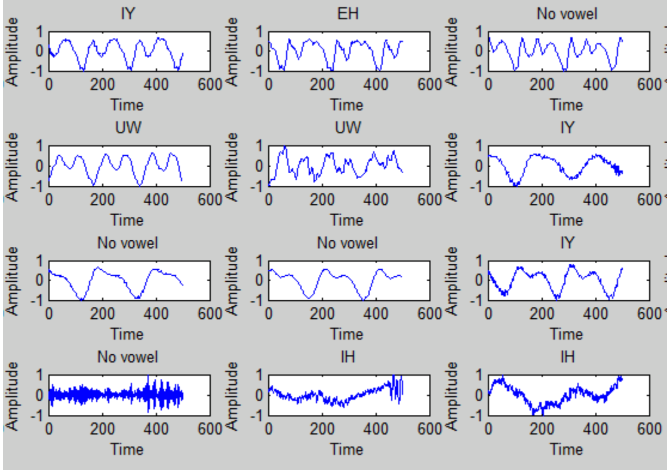

Fig. 7. Normalized Signals of the Vowel Identified of Korean Speaker

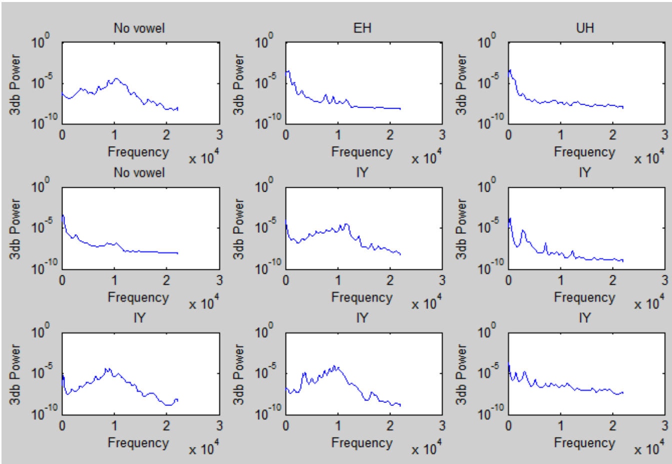

Fig. 8. Stress Intensity Graph of the Vowels Identified of Korean Speaker

- Standard Hindi speakers have a difficulty in differentiating between /v/ and /w/.
- Indian speakers have difficulty in pronouncing the word <our>, saying [aː(r)] rather than [aʊə(r)]. They tend to miss /u/ in their lingual.
- Indian speakers tend to unaspirate the voiceless plosives /p/, /t/, /k/ . The 'h' (exhalation) is generally missing.
- Indian speakers tend to allophonically change the /s/ preceding alveolar /t/ to [ʃ] (<stop>/stɒp/ → /ʃʈap/).
- Indian speakers tend to used /z/ and /dʒ/ interchangeably.
- Both /θ/and /ð/ are pronounced like /t/. Indians tend to replace /pʰ/ with /f/.

3) Japanese Speakers

- For Japanese speakers, slicing phonetic information remains, for example, ("Su" becoming "s"). In this case the small fragment of phonetic sound ("u" sound) is missing, leaving behind awkward vocal sounds.
- In Japanese [v V] /vʌ/ sound closer to the actual pronunciation of [w a] /wa/

  ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

- English has a variety of fricatives and affricates which are more widely distributed than in Japanese [5]. The Japanese consonantal system does not have /f/, /v/, /θ /, /ð/, / ʃ /, / ʒ /, /ʧ/, and /ʤ/.
- Similar to Chinese speakers, the Japanese speakers tend to say the /r/ sound as /l/.
- For Japanese speakers, the time it takes to say a sentence depends on how many syllables in the sentence, but not on how many stressed syllables, as it should be spoken properly [5]. This affects the way that Japanese speakers speak English.
- In Japanese, /ʃ/ /ʒ/ and /ʧ/ /ʤ/ do not usually appear as distinct phonemes. When /s/ /z/ and /t/ /d/ appear before the vowels /I/ and /U/, Japanese speakers tend to allophonically pronounce /ʃ/ /ʒ/ and /ʧ/ /ʤ/ instead.

4) Korean Speakers

- Japanese and Korean seem like allophones and thus have a pitch difference. Japanese tend to speak /t/ less sharply as Koreans.
- Some English vowels are not available in Korean, such as /i/ /v/ / æ/ /ej/ /ow/ /ɔ/ /ə/ / ʌ/ /a/. Thus the pronunciation of Koreans becomes unclear in English.
- Korean speakers tend to pronounce /m, n/ without the nasal effect.
- The phoneme /ŋ/ appears frequently between vowels.
- Koreans tend to say /p/ as /b/, /t/ as /d/, / tɕ/ as /dʑ/, and /k/ as /g/.
- Koreans tend to pronounce /n/ and /l/ similar to /l, hence both /nl/ and /ln/ are pronounced as [lː].

F. Collection of Speech Corpus

To execute the project the most vital part of the software was the selection of speech corpus. To get reliable results having a vast range of speech samples is extremely important. And more the range the standardization of the samples actually contributes positively to the results. These samples can later be used for testing of other speech software and in conducting linguistic studies. These samples helped in comparing the similarities and differences in the above mentioned analysis of languages. The collection of samples was saved as .wav files. The methodology of sample collection is given below:

1. The sentence chosen for the analysis was 'It's a Peaceful Existence'. The reason this sentence was chosen was because it consisted of very distinctive vowels which could certainly help in confirming the vowel identification method. Also in addition it has alphabets which can be used as distinguishing factors for different speaker groups.
2. Forty students from the National Univeristy of Singapore, ten from each of the four nationalities (China, India, Japan, Korea), were recruited to produce speech samples.
3. All speech samples were checked for discrepancies, and were saved as .wav files and filed in folders according to the native language of the speaker.
4. The sample locations can be stored in an SQL database, and linked and accessed by the codes.

The above analysis can be used for several applications such as speech recognition, linguistics research, and studying ESL. It can also help understand how the speaker's own native language affect their pronunciations, and therefore, address their issues more directly.

III. Conclusion and Future work

In this paper, we have presented the results of using MATLAB's Speech Analysis Toolbox for speech feature extraction and data visualization. There are two main areas in this study. In the first part of this work, the formant frequency algorithm was implemented to identify vowels and a success rate of 57.3% was achieved. We anticipate better results with more speech samples in the future. In the second part, we performed a phonological analysis of four Asian languages: Indian, Chinese, Korean, and Japanese. We have provided a general cover graph for every feature of each of the languages, and demonstrated the differences and similarities of the four languages.

This work can be extended in a number of areas. Firstly, more speech features such as Perceptual Linear Prediction (PLP), Relative Spectral PLP can be implemented. Secondly, more languages can be included in the study. Thirdly, automatic recognition systems can be developed based on feature extraction for different purposes. For example, an accent recognition system can be developed using speech extraction and language analysis. Finally, visualization can be improved so the results can be understood better and easier.

References

[1] A. Biem, S. Katagari, and B. H. Juang, "Discriminatiove Feature Extraction for Speech Recognition," in *Proc. 1993 IEEE-SP Workshop*, pp. 392-401.

[2] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Speech Processing*, 1st ed. Pearson, 2010.

[3] B. Prica and S. Ilić, "Recognition of Vowels in Continuous Speech by Using Formants," *Facta universitatis-series: Electronics and Energetics*, vol. 23, no. 3, 2010, pp. 379-393.

[4] K. Brown, *Enclopedia of Language and Linguistics*, 2nd ed. Elsevier, 2010.

[5] K. Ohata, "Phonological Differences between Japanese and English: Several Potentially Problematic Areas of Pronunciation for Japanese ESL/EFL Learners," *Asian EFL J.* vol. 6, no. 4, 2004, pp. 1-19

[6] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. Hoboken, NJ: John Wiley & Sons, 2006.

[7] D. O'shaughnessy, *Speech Communication: Human and Machine*. India:University Press, 1987.

[8] L. Deng and D. O'shaughnessy, *Speech Processing: a Dynamic and Optimization-Oriented Approach*. New York, NY: Marcel Dekker Inc, 2003.

[9] J. Baker, L. Deng, J. Glass, S. Khudanpur, C. H. Lee, N. Morgan, and D. O'Shaughnessy, "Developments and Directions in Speech Recognition and Understanding, Part 1," *Signal Processing Magazine*, vol. 26, no. 3, pp. 75-80, 2009.

[10] Encyclopedia Britannica. (2014). *Pitch* [Online]. Available: http://www.britannica.com/EBchecked/topic/1357164/pitch

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

[11] M. P. Kesarkar, “Feature Extraction for speech Recognition,” M. Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay, Nov. 2003.

[12] B. S. Atal and S. L. Hanauer, “Speech Analysis and Synthesis by Linear Prediction of the Speech Wave,” *J. Acoustical Society of America*, vol. 50, no. 2B, 1971, pp. 637-655.

[13] T. L. Nwe, S. W. Foo, and C. R. De Silva, “Detection of Stress and Emotion in Speech Using Traditional and FFT Based Log Energy Features,” in *Proc. 4th Int. Conf. Information, Communications and signal Processing*, Singapore, 2003, pp. 1619-1623.

[14] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice Hall, 1999.

[15] Punskaya, E, *Basics of Digital Filters* [Online]. Available: http://freebooks6.org/3f3-4-basics-of-digital-filters-university-of-cambridge-w7878/

[16] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[17] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, 1997.

[18] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Urbana: the University of Illinois,1964.

[19] U. Zölzer, *Digital Audio Signal Processing*. John Wiley and Sons, 2008.

[20] D. P. W. Ellis. (2008, October 28). *An Introduction to Signal Processing for Speech* [Online]. Available: http://www.ee.columbia.edu/~dpwe/pubs/Ellis10-introspeech.pdf

[21] J. Cernocky and V. Hubeika. (2009), *Speech Signal Processing – introduction*. [Online]. Available: http://www.fit.vutbr.cz/~ihubeika/ZRE/lect/01_prog_intro_2008-09_en.pdf

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731