World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:6, No:3, 2012

Accent Identification by Clustering and Scoring Formants

Dejan Stantic, Jun Jo

Abstract—There have been significant improvements in automatic voice recognition technology. However, existing systems still face difficulties, particularly when used by non-native speakers with accents. In this paper we address a problem of identifying the English accented speech of speakers from different backgrounds. Once an accent is identified the speech recognition software can utilise training set from appropriate accent and therefore improve the efficiency and accuracy of the speech recognition system. We introduced the Q factor, which is defined by the sum of relationships between frequencies of the formants. Four different accents were considered and experimented for this research. A scoring method was introduced in order to effectively analyse accents. The proposed concept indicates that the accent could be identified by analysing their formants.

Keywords—Accent Identification, Formants, Q Factor.

I. Introduction

Sound is a vibration that propagates through the air or other media. These vibrations are usually caused by vibration of objects such as musical instruments, speakers, or in case of speech, vocal chords. Being able to detect and process these vibrations has tremendous evolutionary advantage for humans and other species. However to process and understand these vibrations is a challenge for computer systems and Artificial Intelligence.

First step in computer speech recognition is the process of digitising the speech signal and transforming it into a set of useful features at a fixed rate. These measurements are then used to search for the most likely word candidate in the database, making use of constraints imposed by the acoustic, lexical, and language models. Throughout this process, training data are used to determine the values of the model parameters.

The majority of speech recognition systems require training by the speaker, where the user provides samples of his or her speech. However there are other systems which are speaker-independent, but the accuracy of those systems are lower. Recognition is generally more difficult when vocabularies are large or have many similar-sounding words and when the dialect or accent of the speaker is different than one who trained the system.

Dialect can be simply described as: 'a variation of a given language spoken in a particular place or by a particular group of people'. While it is considered that a person has an accent when s/he does not sound like a native speaker. Accents usually come from the articulation habits of the speaker in his/her own native language. It has been shown that a speaker's first language affects production and perception of English as a second language and that these effects are experience-dependent [6]. While dialect and accent do not cause that much trouble in communication between humans, it causes a lot of problems for speech recognition systems.

It has been shown that the phoneme, which is the smallest segmental unit of sound, is perceived to have important function by speakers of the specific language or dialect [5]. Also, distinctive patterns arise from differences in vowels and consonants as well as the stress patterns, rhythm, and intonation. These features can vary in the way that they are articulated (e.g. formants, stress patterns, etc).

In this work we address the problem in identifying the speaker's accent based on pronunciation of a short phrase. More specifically, the intention is to identify which acoustic features of the pronunciation play the most important role in identifying the accent. From online available resource we acquired speech of the same text of four commonly present English accents in Australia and analysed which acoustic feature play the most important role in identifying particular accents specifically by pronunciation of particular vowels. We propose a scoring method and in experimental evaluation, we show that our scoring method can prove high probability of identifying accents.

This paper is organised as follows: In the next section we provide some necessary background information related to sound, formants, and short term Fourier transformation. In Section 3 we provide a brief literature review on existing works that has been done in this field. In section 4 we present our scoring method methodology. In section 5 we explain the experimental process involved including the data gathering, analysis and results. Finally, in section 6 we conclude the paper.

II. Background

Understanding the characteristics of a speaker's accent can have tremendous value in better receiving the nature of the speaker's language of origin. However applying their native speech to the English language can prove to be difficult due to the rules of the English language. A common case is the pronunciation of words and certain linguistical elements like stress, rhythm and intonation. The technical side of this research involved the understanding of visual representation of a signal and applying mathematics to calculate the features of the signals. Once calculated, a figure can be produced to further make the investigation more plausible. The following

Dejan Stantic is with School of Information and Communication Technology, Griffith University, Gold Coast, Australia.
Jun Jo is with School of Information and Communication Technology, Griffith University, Gold Coast, Australia.

sections briefly provides information to better understand the content of this research.

A. Sound

Sound is vibration that propagates through the air and other medium. The vibration of the air are caused usually by the vibration of other objects, such as vocal chords, musical instruments, speakers, and so on. A sound has a pitch, which is determined by the frequency with which the pattern repeats itself. This repeating pattern of movements distinguishes sounds from noises. Such a regular movement is described by a sine wave which has a particular frequency.

Being able to detect and process these vibrations has tremendous evolutionary advantage for humans and other species. Our ears analyse sound vibration of different frequencies which are extracted from the sound and sends them to different nerve endings. However, to process and understand these vibrations is a challenge for computer systems and Artificial Intelligence.

B. Formants

The term *formant* refers to spectral peaks in the harmonic spectrum of a complex sound [5]. Formants in the sound of the human voice are particularly important because they are essential components to distinguish vowel sounds. Specifically, there are differences in the first three formant frequencies, which play the most important role, as it can be seen in Figure 1 (This figure is a screen shot of the application which was developed as part of this work).

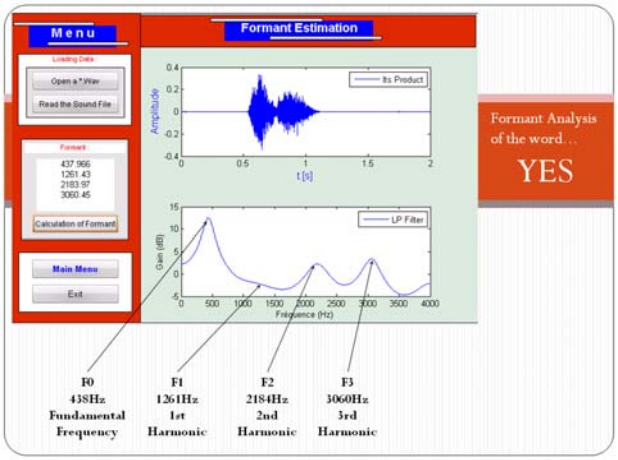

Fig. 1. Formants

Fundamental frequency (F0) is the frequency of repetition of the periodic waveform of the voiced speech signal, corresponding closely to our perception of the pitch of the speech.

C. Short-term Fourier transforms - STFT

Many methods for speech signal analysis start with Fourier transform that gives complete description of signal components in the frequency domain. Very broadly speaking, the Fourier transform is a systematic way to decompose generic functions into a superposition of 'symmetric' functions. These symmetric functions are usually quite explicit (such as a trigonometric function sin(nx) or cos(nx)) and are often associated with physical concepts such as frequency or energy [7].

The short-term Fourier transform (STFT) is used to determine the sinusoidal frequency and phase content of local sections of a signal as it changes over time. For Continuous-time STFT the function to be transformed is multiplied by a window function which is zero for only a short period of time, as it can be seen in Equation 1. The Fourier transform of the resulting signal is taken as the window along the time axis, resulting in a two-dimensional representation of the signal. Mathematically, it can be represented as:

$$STFT\{x(t)\} \equiv X(\tau,\omega) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-j\omega t}\,dt \quad (1)$$

While for discrete-time STFT the data to be transformed could be broken up into frames. Each frame is Fourier transformed, and the complex result is added to a matrix, which records magnitude and phase for each point in time and frequency [15]. This can be expressed as:

$$STFT\{x[n]\} \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \quad (2)$$

with signal x[n] and window $\omega[n]$. In this case, m is discrete and ω is continuous. The magnitude squared of the STFT yields the spectrogram of the function. When the STFT is used for spectral analysis, the $\{x_n\}$, sequence usually represents a finite set of uniformly-spaced time-samples of some signal x(t), where t represents time.

$$spectrogram\{x(t)\} \equiv |X(\tau,\omega)|^2 \quad (3)$$

Many methods for speech signal analysis start with Fourier transform, that gives complete description of signal components in the frequency domain.

III. Related Work

There has been a significant improvement in speech recognition system technology and there are a lot of commercial software (Microsoft Dictation) and freeware (for example, e-Speaking), which in some cases with high accuracy can perform speech recognition. This success can be attributed to the significant attention that this field has received from researchers. Work was mostly concentrated on methods of fast learning as well as identifying which component of the speech plays the most important role for identifying speech.

One of the first work on automatic accent identification dates back to 1986 when McDermott found that a variety of phonological factors influenced listener's judgments of accent, as well as listener background and exposure to foreign languages [11]. Subsequent studies have considered the role of

pronunciation, such as vowel quality [12] in affecting listener perception of foreign accent.

While there are not enough investigations devoted to the accent identification, significant attention was given to the identification of different language dialects, for example, to identify Brazilian Portuguese from European Portuguese [4] and British English from American English [16].

In the literature, attention was also given to identify which specific parts of the speech play an important role in identifying dialect. The acoustic characteristics of the vowels have been considered to play the most important role in identifying dialect, [4], as well as, vowel pronunciation [8].

Another approach has attempted the acoustic-phonetic structure of English regional accents. The modelling of intonation and duration correlates of dialect as the modelling of formants [2], [1]. For example, Australian vowels have longer duration than those of British and American dialect while Australian consonants have shorter duration than those of British and American dialect [16].

It has been reported that every vowel has the maximum frequency and that frequency depends on gender. For example, the median of the 140 cases ceilings for the female was 5450 Hz, and the median of the 140 optimal ceilings for the male was 4595 Hz [4].

It has been identified that the energy, formants and fundamental frequency are the most discriminative features for identifying specifically people with a Cantonese accent. One author stated but has not proved that the same is the case for the other Asian language accents [9].

The dominant recognition concepts use hidden Markov models (HMM). An HMM is a stochastic model, the generation of the underlying phoneme string and the frame-by-frame, surface acoustic realisations are both represented probabilistically as the Markov processes. Neural networks have also been used to estimate the frame based scores, which then can be integrated into HMM-based system architectures [10]. Statistical methods like cluster analysis and multidimensional scaling have successfully been used to determine dialect.

Other approaches mentioned in the literature identify accent parameters which are based on: utilisation of Multiple Livelihood Linear Regression (MLLR) [13], phone N-gram [18], prosodic [14], and word Ngram features [3]. However, these approaches showed to be only applied to specific languages.

Also, it has been proposed to consider relationships between F1, F2 and F3 formants by calculation speaker's k-factor as a ratio of the average F3 value for vowel /a/ (in word "Bark") for one specific speaker divided by the average for all speakers [17].

IV. Scoring Method

In our work we utilise the Fast Fourier Transformation to find the frequency spectrum. As it can been seen in Figure 2 after Fourier transformation, a Linear Predictive Coding (LPC) filter was used to clearly obtain peaks in the frequency domain and identify each formant.

We define the term harmonicity factors:

$$H_{10} = F_1/F_0$$

$$H_{21} = F_2/F_1$$

$$H_{32} = F_3/F_2$$

These factors basically represent numbers defining relative positions of the formants and their relationship on the frequency scale. We identified that this relationship of formants is quite unique and consistent for different accents and for different vowels.

We define the Q factor of the speech as a sum of harmonicities:

$$Q = H_{10} + H_{21} + H_{32}$$

We intend to use Q factor to investigate it's uniqueness for specific accents.

V. Experiment

In the voice signal we isolated the fundamental, first, second and third *Formant*. By using different speakers from different backgrounds, this work looked into which formants, their relationship and position on the frequency scale and other parameters, are crucial for different accents.

The Fourier transform was used to identify spectral analysis. The use of spectral analysis in sound and signal processing gives visual representation of a signal that can further help in identifying key features within the signal. A frequency spectrum gives the frequency on x-axis and amplitude on y-axis of the sound. This is useful for finding certain peaks within the signal and to identify formants. Clustering methods was used for analysing differences in the pronunciation of the vowel in the words.

While trying to reduce the editing and retain the nature of the sound, we opted not to normalise the sound, as our early tests showed no significant difference between the original and normalised sound. Filtering the noise was considered early on, but didn't show dramatic difference between the filtered and non-filtered sound so therefore it was disregarded.

A. Data gathering

Due to the difficulty to record a variety of accents and to find a significant number of people from the same background and same accents, we decided to use suitable recordings available in an online archive. We found sound files on the Speech Accent Archive web site (http://accent.gmu.edu/). On this web site a variety of accents can be found where all the speakers are asked to read several sentences.

The sentences contained certain words that made strong emphasis on the vowels. This was useful for us as we wanted to investigate characteristics of different accents and its influence to the vowel pronunciation. The sentence we chose is as follows:

"Ask her to bring these things with her from the store."

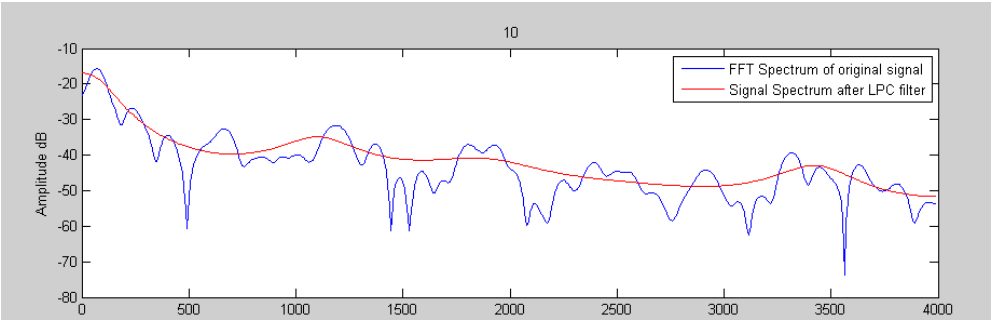

Fig. 2. LPC Filtering

We have chosen one sentence with strong emphasis on the vowels in order to investigate how on average different words (vowels) within the sentence influence distribution of formants for different accents. We decided to investigate several English accents, which are typical representatives of accents found in Australia.

In making the decision which speakers to consider, we have taken into consideration the age of the speakers, as it has been shown that older age people have a stronger accent. We have taken into consideration the onset of learning English as well as disregarding candidates who have spent more than 10 years in English speaking countries.

The following English-speaking accents are considered:

- Chinese
- Indian
- Korean
- Croatian

We have conducted relatively a small scale experiment consisting of about 15 speakers from each English-speaking accent. Despite the relatively small scale we were able to draw conclusions and to identify patterns.

In order to have consistency we have chosen the same gender, in our case we have chosen the 'male' gender. The voice range of females is not as consistent as males, therefore having to neglect female candidates for this research. Despite using online available source data, it contained the consistency we were looking for. Keeping in mind the nature of this work, to recognise accents from speech mostly obtained over the telephone, we decided not to alter the original sound files. As most recordings contained some sort of background noise and ambience, the use of normalisation and filtering was bypassed to correspond to everyday real life situations.

The process of gathering the sound files involved extracting the MOV files. Once all the necessary files were extracted, they needed to be converted into WAV audio only format. Using the RealPlayer application, we converted the MOV files into WAV formant with a bit rate of 16 by 44.1Khz sample rate. Once converted, it needed to be down sampled to the sample rate of 8000Hz while keeping the bit rate the same for the use of the analysis application, commonly used for phonetic and linguistic analysis, which we developed in Matlab. Reason for dropping the sample rate is that we are only looking at the frequency range of the human voice which ranges from 60 to 7000Hz, plus we intend to uncover the fundamental, first, second and third formant frequencies.

$Accent$	$"QFactor"$
$Korean1$	6.68
$Korean2$	6.55
$Korean3$	6.49
$\dots$	$\dots$
$Korean15$	6.33
$Average$	6.57

TABLE I
Q FACTOR OF KOREAN ACCENT

This maximum frequency of the sampling rate is chosen to enable to capture the third formant, which is usually under 4000Hz. However it is important to mention that in order to squeeze more phone lines into a single wire, the telephone companies limits the frequency bandwidth of phone calls between, roughly 350Hz and 3,500Hz. Since normal human hearing ranges from 20Hz to 20,000Hz, the limited frequency response in a phone call gives it a unique sound.

Out of all the recordings we decided to look more closely into the sentence one: "Ask her to bring these things with her from the store". Finding the formants of each speaker involved going through each preprocessed recording. The formant data was automatically recorded out from the application developed in Matlab into Microsoft Excel spreadsheets for further analysis.

B. Results and Analysis

We looked into frequency distribution and formants positions of the whole sentence. Specifically, we identified that only the relative relationship of formant F1 and F2 clustered very well together for the same accent, as can be seen in Figure 3.

Also, while looking into the data and relationship of formants, we identified that for different accents different relationship of formants are more specific for particular vowels.

This difference is even more evident when a whole sentence is taken into consideration Figure 4, which provides evidence that the accent is possible more easily to identify if a whole sentence is considered. This is in line with when humans are listening to English accent and trying to identify it. Basically

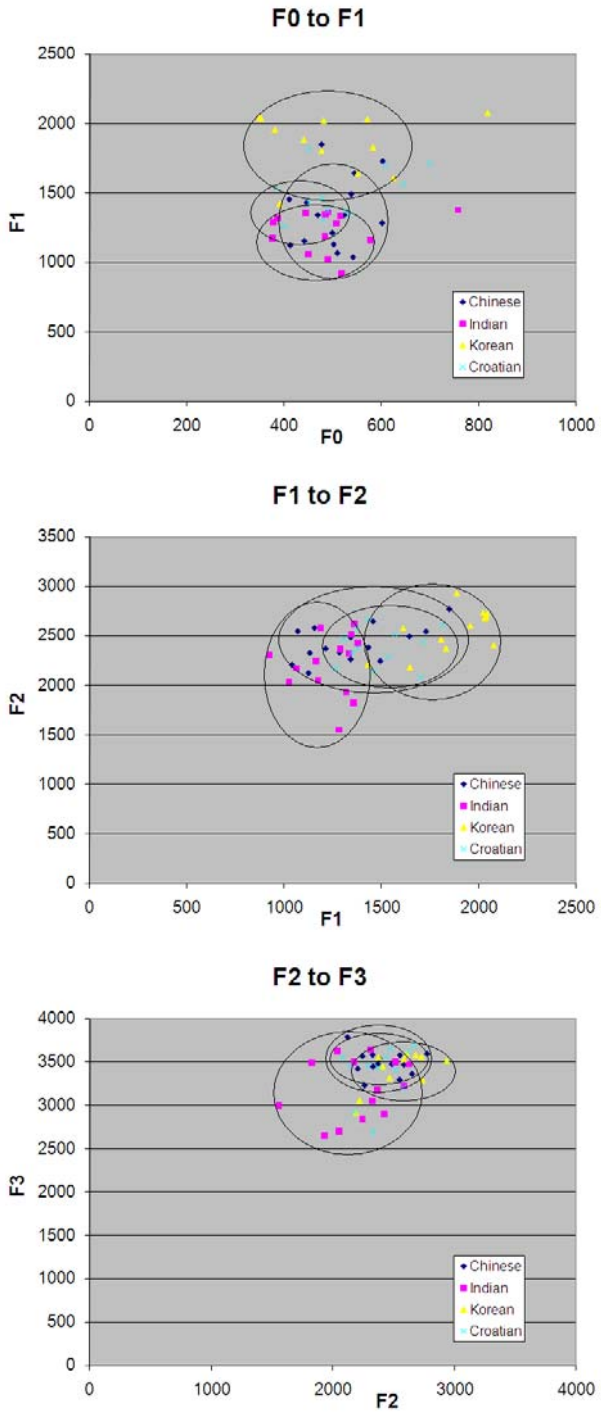

Fig. 3. Clustering of formants F0 to F1, F1 to F2 and F2 to F3.

different vowels in different accents have more dominant formant and characteristics that can be used to identify accent. Obviously, this difference influences the Q factor.

As based only on clustering of different relationships of formants, it is not possible with high probability to identify the accent. We looked into the sum of relationships and distances of formants for different accents, as explained in section 4., which we termed Q factor. As it can be seen in Table I this Q factor is relatively consistent for the same accent as it depends basically on a specific vowel.

Accent	*"QFactor"*
Chinese	6.01
Indian	5.88
Korean	6.57
Croatian	6.02

TABLE II
Q FACTOR OF PARTICULAR ACCENT FOR THE SENTENCE.

Table II shows some data for the Q factor for the different accents for the sentence. It can be seen that this Q factor is different for each accent and is dependent on a specific vowels within the sentence. It seems that Indian and Korean accents are far more distinguished then the other two accents. This shows that the Q Factor can be applied to the recorded data and calculated to give a reading to help identify the appropriate accent if ever wanting to incorporate into a speech recognition system.

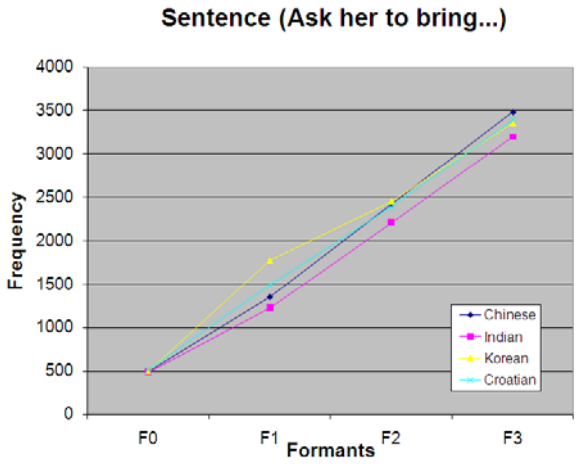

Fig. 4. Distribution of Formants for the Sentence.

VI. Conclusion

In this work we addressed the problem of automatically identifying the English accent based on pronunciation of a short phrase. If the accent can be identified at the beginning of the usage of the automated customer service then a appropriate training set can be plugged in the speech recognition system, which in turn will ensure better accuracy of speech recognition and better efficiency. Specifically, in this work we introduced the Q factor and showed that this factor can be used to identify certain English accents with high probability.

As for future work, we intend to further investigate the consistency of Q factor and to incorporate it within the speech recognition system in order to investigate how the accuracy and efficiency is affected by applying the training set suggested by the Q factor and clustering method.

REFERENCES

[1] K. Bartkova and D. Jouvet. Automatic detection of foreign accent for automatic speech recognition. *In Proceedings of the International Congress of Phonetic Sciences ICPS07*, pages 2185–2188, 2007.

[2] T. Chen, C. Huang, E. Chang, and J. Wang. Automatic accent identification using Gaussian mixture models. *In Proceedings of the IEEE Workshop on Automatic Speech Recognition*, pages 343–346, 2001.

[3] G. Doddington. Speaker recognition based on idiolectal differences between speakers. *In the Proceedings of the 5th European Conference on Speech Communication and Technology - Eurospeech01, Aalborg, Denmark*, pages 2521–2524, 2001.

[4] Paola Escudero, Paul Boersma, Andreia Schurt Rauber, and Ricardo Bion. A Cross-dialect Acoustic Description of Vowels: Brazilian and European Portuguese. *Journal of the Acoustical Society of America*, 126(3):1379–1393, 2009.

[5] G. Fant. *Acoustic Theory of Speech Production*. Mouton and Co, The Hague, Netherlands, 1960.

[6] James Emil Flege, Ocke-Schwen Bohn, and Sunyoung Jang. Effects of experience on non-native speakers production and perception of English vowels. *Journal of Phonetics*, 5(1):437–470, 1997.

[7] M. Greitans. Adaptive STFT-like Time-Frequency analysis from arbitrary distributed signal samples. *International Workshop on Sampling Theory and Application*, 2005.

[8] Therese Leinonen. Factor analysis of vowel pronunciation in swedish dialects. *International Journal of Humanities and Arts Computing*, 2(1):189–204, 2009.

[9] Gina Levow. Investigating pitch accent recognition in non-native speech. *In the Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, Singapore*, pages 269–272, 2009.

[10] S. Matsunaga, A. Ogawa, Y. Yamaguchi, and A. Imamura. Non-native English speech recognition using bilingual English lexicon and acoustic models. *In the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP03*, pages 340–343, 2003.

[11] W. C. McDermott. *The Scalability of Degrees of Foreign Accent*. PhD thesis. Cornell University, 1986.

[12] M. J. Munro, T. M. Derwing, and J. E. Flege. Canadians in Alabama: A perceptual study of dialect acquisition in adults. *Studies in Second Language Acquisition*, 27:385–403, 1999.

[13] K.J. Preacher, P.J. Curran, and D.J. Bauer. Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis. *Journal of Educational and Behavioral Statistics*, 31(4):437–448, 2006.

[14] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. Modeling prosodic feature sequences for speaker recognition. *Speech Communication, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation*, 46(2):455–472, 2005.

[15] Kamil Wojcicki, Mitar Milacic, Anthony Stark, James Lyons, and Kuldip Paliwal. Exploiting conjugate symmetry of the short-time fourier spectrum for speech enhancement. 2008.

[16] Qin Yan and Saeed Vaseghi. Modeling and synthesis of English regional accents with pitch and duration correlates. *Computer Speech and Language*, 24:711–725, 2010.

[17] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.Y. Yoon. Accent detection and speech recognition for Shanghai-accented Mandarin. *In the Proceedings of the 9th European Conference on Speech Communication and Technology - Eurospeech05*, pages 217–220, 2005.

[18] M. A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling. *In the Proceedings of the Acoustics, Speech, and Signal Processing ICASSP, Adelaide, Australia*, pages 305–308, 1994.

Dejan Stantic Dejan Stantic completed Bachelor of Multimedia with first class Honours. He previously graduated Popular Music and Digital Video Production where he got interested in sound and signal processing. His research interest focuses on signal formant analysis. Currently he is a PhD student where he extended his work on signal analysis to the health domain and electrocardiogram signals.

Jun Jo Dr. Jun Jo was awarded his PhD degree from the University of Sydney in 1994. He has conducted research in various areas including Computer Vision, Robotics and e-learning and has published over 100 refereed publications. Dr. Jo is currently taking the positions of the Director General of International Robot Olympiad Committee (IROC) and the President of Australian Robotics Association (ARA).