Signal Processing Techniques for Language Identification Based on the Pitch Contour

Stephen D. Bier*, Catherine I. Watson*, Margaret Maclagan†, Jeanette King‡ Ray Harlow§ and Peter Keegan¶
*Department of Electrical and Computer Engineering, The University of Auckland, Auckland, New Zealand
Email: s.bier@auckland.ac.nz, c.watson@auckland.ac.nz
†Department of Communication Disorders, University of Canterbury, Christchurch, New Zealand
Email: margaret.maclagan@canterbury.ac.nz
‡Aotahi School of Māori and Indigenous Studies, University of Canterbury, Christchurch, New Zealand
Email: j.king@canterbury.ac.nz
§Hamilton, New Zealand
Email:rharlow@waikato.ac.nz
¶Faculty of Education, The University of Auckland,Auckland, New Zealand
Email: p.keegan@auckland.ac.nz

***Abstract*—In this paper we present the signal processing techniques used in two different language identification tasks, a perception based task and an automated task. The two languages studied were English and Māori, and the language idenitifcation is done on the "melody" of the speech, i.e. the suprasegmental features rather than on explicit sound recognition. We argue that for language identification based on suprasegmental features it is more useful to perform these studies on explicit prosodic features such as pitch and loudness. New stimuli based on complex harmonic sinusoidal synthesis are presented, these stimuli have been successfuly used in a perceptual based language identification task, showing that only pitch is required to identify the two languages. A method of encoding and analysing the pitch contour with DCT coefficients is presented and able to show significant differences between the c_0 and c_3 coefficients for read excerpts of Māori and English. Finally we show that the DCT coefficients of the pitch contour can be used in an automatic language identification task to distinguish between English and Māori with a classification rate of 71%.**

I. Introduction

Automatic Language Identification (LID) is an important part of automatic speech recognition. It informs what words and grammar to parse for in any given speech segment. Human listeners are generally able to identify languages that they are familiar with, but for automatic LID computers need to be trained with models that allow them to perform such tasks. While it is possible to build models based on full speech recordings, it would be computationally advantageous to reduce the models down to simpler parameters. To investigate what parameters might be useful for automatic LID, it is possible to identify what cues human listeners are able to base their judgments on.

There are in fact two distinct approaches to LID, those done by speech technologists for automatic LID, and those done by linguists to determine what acoustic cues people use to distinguish languages, that is perception studies. In this study we first draw on the work done by linguistics and use that to inform our speech processing approach. To illustrate our approach we use English and Māori as our two languages to distinguish, however the methodology should be able to be applied to other languages.

A. Previous perceptual LID experiments

Many studies have assessed listeners' ability to discriminate between languages using modified speech, in order to measure the effect of prosody on LID. Common modifications include low-pass filtering [1][2][3], use of laryngograph waveforms [3][4][5], resynthesised pulse trains [6][7], LPC filtering [8], and segmental resynthesis [9]. Different modifications leave different aspects of prosodic and linguistic information in the speech samples for listeners to use as cues for language identification. For most cases the modified waveforms are intended to retain varying combinations of pitch, rhythm and intensity while eliminating as much lexical information as possible.

Low-pass filtered speech is one of the simplest modifications used in perceptual LID experiments, as discussed by Komatsu [10]. It has shown various languages to be discriminable in previous studies such as Maclagan et al. [1] and Atkinson [2], but low-pass filtering in particular makes it hard to remove all the lexical content without taking away from the pitch contour. As Ramus and Mehler [9] state, low-pass filtering "does not allow one to know which properties of the signal are eliminated and which are preserved". Thus some of the segmental information may have unintentionally been preserved. Additionally, low-pass filtering does not provide much reduction in parameters for analysis, so does not provide much computational benefit with regard to automatic LID.

Other modifications draw upon the source filter model of speech production, with the aim of getting an indication of the source or glottal pulse, which does contain a lot of prosodic information. The simplest of these approaches is to audibly play electroglottograph waveforms, which while not directly equivalent to the glottal pulse, are still closely related to them. It was shown by Maidment [5] that English and French could be discriminated in this way and by Moftah and Roach [3]

that English and Arabic could be differentiated. Moftah and Roach also showed that there was no significant difference in LID accuracy between laryngograph waveforms and low-pass filtered speech.

There is a range of synthesised stimuli that have been investigated in order to further eliminate any interference from lexical information. The creation of these stimuli begins with calculating the pitch contour using an algorithm such as MOMEL [6] or ESPS [11]. The pitch contour can then be resynthesised as a series of pulses [6] or by the insertion of speech like signals. In the study by Ramus and Mehler [9] the suprasegmental information was retained in the stimuli by converting all the vowels to /a/ and all the consonants to /s/, giving listeners an increased idea of the rhythm of the speech with what they called *sasasa* stimuli. However, Szakay [12] found that when the *sasasa* style of stimuli is applied to Māori, the sound is too far removed from real speech for appropriate LID to be performed.

Many of the LID studies have looked at comparing languages that are considered to be in different rhythm categories such as comparing tonal languages with non tonal ones [10]. When limiting the cues for listeners to pitch and intensity, both of these types of comparison are more likely to yield significant results than comparisons between more similar languages.

B. Justification for pure pitch contour based LID

What is desired in prosody based stimuli for LID tasks is a waveform that contains pitch and intensity without any other information. For many of the stimuli used in previous LID tasks it is debatable whether this is achieved.

As previously pointed out, in addition to only providing minimal parameter reduction, low-pass filtering does not allow one to know exactly what information is preserved and what information is removed. Typical filter cutoff frequencies also tend to be 500 Hz or lower, taking the stimuli below optimal hearing ranges. Figure 1 shows the sound pressure levels in dB required for equivalent volumes at different frequencies in normal listeners [13]. Hearing thresholds drop as frequency increases toward 1000 Hz and are much lower from 500 Hz upwards. Therefore stimuli that contain frequencies above 500 Hz are much easier for listeners to hear in LID tasks.

Estimates of the glottal pulse are more likely to contain only the prosodic detail that is being investigated, and are also likely to contain frequencies above 500 Hz that assist with hearing. Glottal pulse estimates based on electroglottograph recordings, whilst being able to provide a good estimate, are impractical for many LID applications as they require measurements other than just audio, which are often unobtainable. When looking at the perceptual case, the premise that glottal flow waveforms are appropriate for LID tasks is flawed in some respects, as we never hear the glottal flow without the effect of the vocal tract as a filter. Furthermore, when these glottal waveforms are obtained through inverse LPC filtering, it is possible that some lexical information is still preserved in the waveform. As Komatsu [10] points out, LPC filtering leaves some of the higher spectral components behind, resulting in the need to further modify the signal in an effort to remove lexical information.

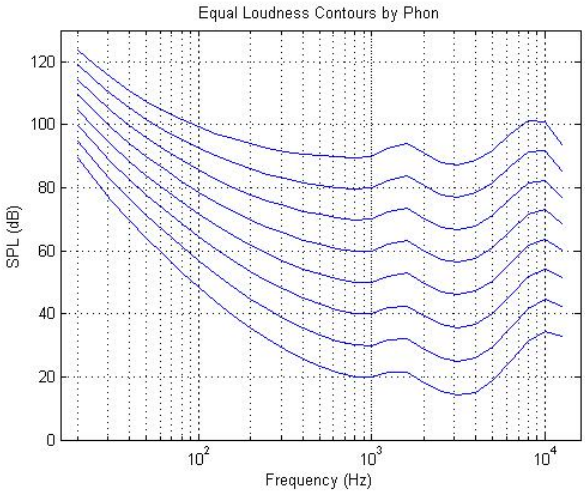

Fig. 1. *Equal loudness thresholds for normal hearing [13]*

With regard to generation of speech-like stimuli, there are a number of issues that can be encountered. Firstly, there is a chance that the phonemes being used are not part of one or more of the languages in a given LID task. Secondly, different vowels have different intrinsic effects on the prosody, so there may be some psychoacoustic interference when one specific vowel is overlayed over all the vowel sounds [14].

For these reasons, it is proposed that a better approach to generating stimuli for LID tasks may be to take the pitch contour of the speech samples being modified and resynthesize it with a complex tone. This signal can then be modulated by the intensity of the original speech sample. If these parameters are sufficient for perception based LID, then they should prove useful for developing an automatic LID model.

II. Stimuli creation

As indicated in section I-B, it is desired to produce LID stimuli that:

- contain both pitch and intensity information
- do not contain any lexical information
- are easily audible
- can be related to speech

For the stimuli discussed in this paper the pitch was calculated using the ESPS algorithm [11] and the intensity was calculated from the root mean square (RMS) energy of the samples. Pitch and intensity contours were produced with EMU (http://emu.sourceforge.net/) and read into R (http://www.r-project.org/) for further processing. The stimuli are able to be generated purely from the pitch and intensity contours of the original speech segments.

The stimuli generated are produced as a sum of harmonically related sinusoids, with the frequency of the lowest sinusoid being defined by the pitch contour of the original

stimuli. In order to produce the sinusoid, the pitch and intensity contours were first interpolated to the desired sampling frequency of the final stimuli (in this case, 22050Hz). The base sinusoid is then generated according to the equation

$$s(t) = A(t) \sin\left(2\pi \int_0^t f_0(\tau)\, d\tau\right) \quad (1)$$

where $A(t)$ is the intensity contour, $f_0(\tau)$ is the pitch contour, τ is instantaneous time and t is the time in general. Being performed numerically in discrete time, Equation 1 becomes

$$s[n] = A[n] \sin\left(2\pi \sum_{k=1}^{n} f_0[k]\right) \quad (2)$$

where k and n are sample numbers, k being the discrete time equivalent of instantaneous time and n being the discrete time equivalent of general time.

This produces a pure tone representation of the pitch contour, modulated by the amplitude. However, pure tones below 500 Hz can be difficult for listeners to hear [13]. A preliminary study was performed using synthetic stimuli composed of a pure sinusoid without any higher harmonics. For all speakers neither the Māori nor English excerpts were correctly identified at a rate significantly better than chance. Based on feedback, it was determined that the softness of the stimuli would have made the language discrimination task difficult.

We wish for listeners to be able to clearly hear and follow changes in the pitch. If a complex tone is used, pitch discrimination can be aided by the first five harmonics [15]. Thus, the first five harmonics were added into the stimuli, with Equation 2 becoming

$$s[n] = A[n] \sum_{j=1}^{5} H_j \sin\left(2\pi \sum_{k=1}^{n} j f_0[k]\right) \quad (3)$$

where H_j represents the proportion of the fundamental and each of the next 4 harmonics. For the stimuli that were generated it was decided to have a gradual roll off of the higher frequency components in order to emulate roll off seen in typical glottal pulse spectrums, so values of 1, 0.5, 0.4, 0.3 and 0.2 were chosen for H_1 through to H_5 respectively. This produced a waveform that looks approximately like a sawtooth, and these stimuli were found to be much easier to hear than pure sine waveforms. The stimuli were normalised and made into wav files. A section of one of the stimuli can be seen in Fig. 2. The top half of the figure shows the approximate sawtooth shape of the resynthesised waveform with changing intensity and the spectrogram shows that the waveforms have energy ranging up to about 1000 Hz, changing with the pitch contour.

In addition to creating stimuli with both pitch and intensity information, stimuli with just pitch were made by holding $A[n]$ constant and stimuli with just intensity information were made by holding $f_0[n]$ constant.

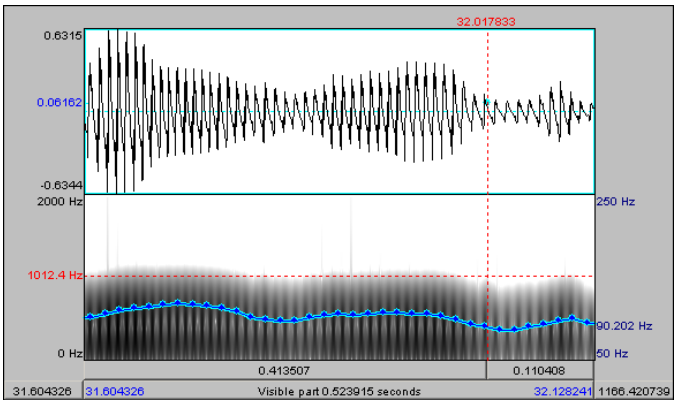

Fig. 2. *Resynthesised stimulus viewed in Praat, showing time domain (top), spectrogram and pitch contour (bottom). Note that pitch contour is superimposed over the spectrogram, but they have seperate y axis scales.*

III. Perceptual LID based on pitch and loudness cues

Following the low-pass filtered LID experiment by Maclagan et al. [1], an LID task with synthetic stimuli was performed in Watson et al. [16] as a continuation of the MAONZE (MAOri and New Zealand English) project [17], an investigation into sound change over time in Māori, the indigenous language of New Zealand. As such, for the purposes of this paper the focus is on the distinction of Māori and English. While this experiment is presented in more detail in Watson et al. [16], a summary is presented here.

For the study in Watson et al. [16] 9 speakers were selected who were representative of the 67 speakers in the wider database. For each speaker two excerpts were used, one in Māori and one in English. These excerpts were each about 15s long and were a subset of the excerpts used in the low-pass task in [1]. The excerpts were taken from conversational speech.

Three different sets of stimuli were then created according to Equation 3:

- just pitch information (P)
- just intensity information (RMS)
- both pitch and intensity and information (PRMS)

The testing procedure followed the same format as the earlier low-pass based LID task in Maclagan et al. [1].

Using these synthetic stimuli, Watson et al. [16] found that stimuli that included pitch information (the P and PRMS conditions) resulted in language identification between English and Māori being significantly greater than chance. The P and PRMS conditions had correct identification rates of 59% and 61% resepectively, slightly lower than the low-pass condition in Maclagan et al. [1] where 68% were correctly identified. This would suggest that some segmental information may have influenced the low-pass results. These results also suggest that the pitch can provide cues that help in identifying Māori and English.

Given that the pitch contour is providing sufficient cues to obtain accurate language identification, quantitative analysis

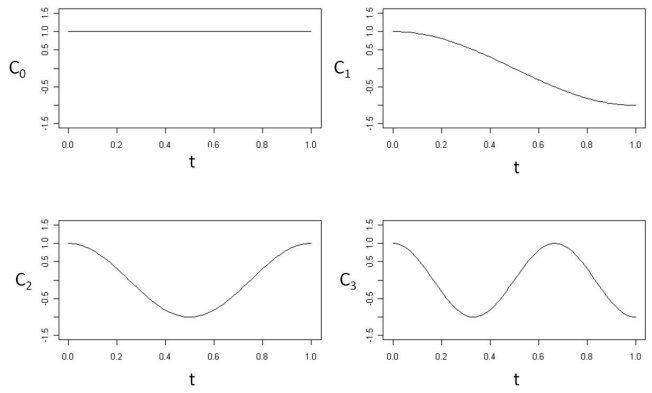

Fig. 3. *Normalised cosines corresponding to the first 4 DCT coefficients c_0 through to c_3*

of these pitch contours is necessary to see what the listeners might be picking up on. A measure such as the discrete cosine transform is able to quantify the shapes of the pitch contours at a phrase level as done by Teutenberg, Watson and Riddle [18].

IV. Discrete Cosine Transform

The discrete cosine transform (DCT) represents a waveform as the sum of a collection of weighted cosines of different frequencies. The frequencies are integer multiples of the frequency corresponding to a period of double the length of the waveform being analysed. The contributions of the first 4 DCT coefficients can be seen in Figure 3, where it can be seen that c_0 corresponds to the mean of the waveform, c_1 gives an indication of overall slope, and higher coefficients continue to add higher frequency components to the shape of the waveform. The DCT coefficients or cosine weights are calculated by Equation 4, where c_k is the kth DCT coefficient, M is the number of samples in the signal, and $f(t)$ is the original signal, which in this case is the pitch contour of a given phrase.

$$c_k = \sum_{t=0}^{M-1} f(t)\, cos\left(\frac{\pi}{M} k \left(t + \frac{1}{2}\right)\right) \tag{4}$$

The signal can be recreated using the inverse DCT, given by Equation 5, where N is the number of coefficients used.

$$f(t) = \frac{1}{2}c_0 + \sum_{k=1}^{N-1} c_k cos\left(\frac{\pi}{M} k \left(t + \frac{1}{2}\right)\right) \tag{5}$$

The number of coefficients used, N, is limited to being no greater than the number of sample points, M. When N is equal to M the original waveform is reconstructed perfectly, but using a lower number of coefficients effectively low-pass filters the signal. By its nature, the DCT normalises for the length of the signal being analysed.

A. DCT analysis of pitch contours

For this analysis, instead of the conversational speech excerpts used by Maclagan et al. [1] and Watson et al. [16], excerpts of read Māori and English passages were divided into their natural phrases, as classified by language grammar rules, and checked based on audible pitch resets of the speakers in the recordings. The analysis of the pitch contour was performed on these natural phrases, at a similar level to the phrase level identified by Teutenberg, Watson and Riddle [18] and Fujisaki [19]. Pitch contours were produced with EMU (http://emu.sourceforge.net/) using the ESPS algorithm and read into R (http://www.r-project.org/) for further processing.

The phrases were split between present day Māori youth and elders. Only male speakers were analysed to maintain consistency with the LID task. A total of 288 phrases were analysed (129 Māori, 159 English) across 13 speakers.

For each phrase, linear interpolation was performed in the unvoiced sections as per Teutenberg, Watson and Riddle [18]. The pitch values were converted to the log frequency domain, as done by Fujisaki [19]. This effectively normalises for different pitched voices, and is necessary as pitch perception is not linear. DCT coefficients were then computed for each phrase. This analysis only looked at the first 5 DCT coefficients, as Teutenberg, Watson and Riddle [18] found that 5 coefficients were sufficient to model the pitch contours with a low RMS error.

B. Results

In addition to language we were also interested in whether the age of the speakers impacted on LID. The reason for this is twofold. Firstly, our earlier studies [17][20] have shown a change in Māori over time, with modern Māori having more features in common with English. Secondly, it is well known that aging has an impact on pitch [21][22].

Therefore a two-way ANOVA was performed for each of the first 5 DCT coefficients (c_0 to c_4), with the coefficients being the dependent variables and language and speaker group being independent variables. There were significant interactions between c_0 and language ($F(1,284)=24.71$, $p<0.001$), c_0 and speaker group ($F(1,284)=8.49$, $p<.01$) and c_0, language and speaker group ($F(1,284)=5.21$, $p<.05$). c_0 relates to the average pitch of each segment. Post hoc Tukey tests showed c_0 to be significantly lower for English than Māori overall ($p<.001$) and significantly lower for elders than youth overall ($p<.01$). When breaking it down by both speaker group and language, c_0 was signicantly lower for elders speaking English than elders speaking Māori ($p<.001$) or youth speaking English ($p<.01$). There was no significant difference in c_0 between the youth speaking Māori or English. This shows that the elders spoke with a lower mean pitch in English than Maori, whereas the youth did not make this distinction. Examining the mean values of c_0 and calculating the corresponding average pitch in Hz gives values of 109.4, 120.2, 126.0, and 126.8 for elders speaking English, youth speaking English, elders speaking Māori, and youth speaking Māori respectively.

There were also significant interactions between c_3 and language ($F(1,284)=13.89$, $p<0.001$), c_3 and speaker group ($F(1,284)=4.08$, $p<.05$) and c_3, language and speaker group ($F(1,284)=10.05$, $p<.01$). Post hoc Tukey tests showed c_3 to be significantly higher for Māori than English overall ($p<.001$) and significantly higher for elders than youth overall ($p<.05$). When broken down by speaker group and language, c_3 was found to be significantly higher for elders speaking Māori than for elders speaking English ($p<.001$) or for youth speaking Māori ($p<.01$). The mean c_3 values were -0.017, -0.0095, 0.049, and -0.00021 for elders speaking English, youth speaking English, elders speaking Māori, and youth speaking Māori respectively. Notably the mean value for elders speaking Māori is positive, whereas the other mean values are negative. A positive c_3 coefficient corresponds to a down-up-down shape within the pitch contour as seen in Figure 3, whereas a negative 4th coefficient corresponds to an up-down-up shape.

There were no significant interactions apparent in the other DCT coefficients.

C. Discussion

From the ANOVA's we can see that the elders do have certain distinctions between their Māori and their English. The most obvious distinction is the difference in average pitch, with their Māori being about 15 Hz higher than their English in the read excerpts. This certainly could be a contributing factor toward correct LID based on pitch contours. For these phrases it is likely that the trend of English being at a lower pitch than Māori overall is caused by the difference within the elders, as the youth do not display a significantly higher pitch for their Māori than their English. That said, they did still have a marginally lower mean pitch for their Māori phrases than their English, so a larger study with greater statistical strength might be able to show this as significant.

The differences in c_3 are perhaps more interesting in regard to the phonetics, as they actually give us an indication that there are different shapes of pitch contour going on between the languages. Although once again the overall trend is probably largely due to the trend for the Māori elders. The lack of such distinctions in the youth is consistent with the idea that Māori is becoming more similar to English over time.

The fact that the polarity of c_3 was reversed for the elders speaking Māori has interesting implications regarding the overall shape of the pitch contours. However, while this DCT analysis can provide a rough measure of the shape of the pitch contours, more detailed phonetic analysis of the phrases should be performed in order to further investigate the pitch contours of the two languages. Similar analysis should also be done on the conversational excerpts, to see if the results are consistent with the results for read excerpts.

V. Automated LID based on DCT coefficients

As there were some significant language differences within the DCT coefficients analysed, a preliminary model for automatic LID based on the first 5 DCT coefficients was made using quadratic discriminant analysis, with the methods used by Harrington [23]. As this was just a proof of concept, a closed classification model was used, which was trained on all the phrases used in the DCT analysis with varying numbers of DCT coefficients. The overall LID accuracy for distinguishing between Māori and English was then calculated.

When the model was trained purely on c_0 it acheived an accuracy of 61.8%, which is about the same as the accuracy acheived by the human listeners in the perceptual LID experiment. When c_3 was also added to the model the accuracy increased to 66.7%. Using the first 5 coefficients further improved the accuracy to 71.2%, which begins to surpass the results from the low-pass filtered LID task in [1].

Given that the DCT analysis above showed differences between the two speaker groups, classification models were also made within the speaker groups. Training based on c_0 produced a higher accuracy for the elders (68.5%) than the youth (57.1%). This is consistent with the finding that the elders make a larger pitch distinction between language than the youth. Using c_0 and c_3 the accuracy was also greater for the elders (74.8%) than the youth (64.6%), and this trend continued when using the first 5 coefficients (elders 77.2%, youth 70.2%). This continues to support the idea that the elders are making greater distinction between languages than the youth, which in turn supports the idea that Māori is becoming more similar to English over time.

These results indicate that a simple DCT model of the pitch contour is able to achieve similar or greater LID accuracy than prosody based perceptual LID tasks. DCT based models of the pitch contour are also able to be made using relatively few data points per phrase, saving on computational cost. As such, it is worth further developing DCT based automatic LID.

VI. Conclusions and future use

The use of resynthesised pitch contours has been shown to be effective at conveying prosodic information. When producing resynthesised pitch contours, the addition of higher harmonics improves audibility and thus improves LID performance. As such, in LID experiments using these complex harmonic tones, listeners were able to correctly identify Māori and New Zealand English at a rate better than chance when the stimuli provided pitch information. This indicates that the pitch of speech provides information that may be useful for automatic LID.

DCT models of the pitch allow quantitative comparisons to be made between pitch contours of phrases from different languages. These comparisons are able to show significant differences between read excerpts of Māori and English, allowing the DCT to be useful for the developement of automatic LID. In the future this method of analysis should be extended to other languages, as different parameters may be needed to identify some languages. It is important to continue investigating prosodic and linguistic features of speech so that automatic LID systems can be developed that are capable of handling the changes within languages over time.

ACKNOWLEDGMENTS

We would like to thank the Marsden Fund of the Royal Society of New Zealand and the University of Canterbury for providing funding to support this research.

REFERENCES

[1] M. Maclagan, C. I. Watson, J. King, R. Harlow, L. Thompson, and P. Keegan, "Investigating changes in the rhythm of Māori over time," in *10th Annual Conference of the International Speech Communication Association: Interspeech 2009, Brighton, September 6-10*, 2009, pp. 1531–1534.

[2] K. Atkinson, "Language identification from nonsegmental cues," *The Journal of the Acoustical Society of America*, vol. 44, p. 378, 1968.

[3] A. Moftah and P. Roach, "Language recognition form distorted speech: Comparison of techniques," *Journal of the International Phonetic Association 18*, pp. 50–52, 1988.

[4] J. Maidment, *Voice fundamental frequency characteristics as language differentiators*. London: University College, 1976.

[5] J. A. Maidment, "Language recognition and prosody: Further evidence," in *Speech, hearing and language: work in progress 1*, 1983, pp. 133–141.

[6] M. Komatsu, T. Arai, and T. Sugawara, "Perceptual discrimination of prosodic types and their preliminary acoustic analysis," in *Proceedings of Interspeech 2004*, 2004, pp. 3045–3048.

[7] M. Barkat, J. Ohala, and F. Pellegrino, "Prosody as a distinctive feature for the discrimination of Arabic dialects," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[8] J. Navratil, "Spoken language recognition-a step toward multilinguality in speech processing," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 6, pp. 678–685, 2001.

[9] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: A study based on speech resynthesis," *Journal of the Acoustical Society of America*, 1999.

[10] M. Komatsu, "Reviewing Human Language Identification," *Speaker classification II*, pp. 206–228, 2007.

[11] W. Kleijn and K. Paliwal, *Speech coding and synthesis*. New York, NY, USA: Elsevier Science Inc., 1995.

[12] A. Szakay, *Ethnic Dialect Identification in New Zealand: The Role of Prosodic Cues.* VDM Verlag, 2008.

[13] B. ISO, "226: 2003: Acoustics - Normal equal loudness-level contours," *International Organization for Standardization*, 2003.

[14] J. Clark and C. Yallop, *An introduction to phonetics and phonology*. Malden, MA: Wiley-Blackwell, 1995, pp. 331–339.

[15] B. Moore and R. Peters, "Pitch discrimination and phase sensitivity in young and elderly subjects and its relationship to frequency selectivity," *The Journal of the Acoustical Society of America*, vol. 91, p. 2881, 1992.

[16] C. I. Watson, J. King, S. Bier, M. Maclagan, R. Harlow, L. Thompson, and P. Keegan, "Prosodic clues in language recognition: How much information do listeners need to identify Māori and English?" *Te Reo*, vol. 54, pp. 83–112, 2011.

[17] R. Harlow, P. Keegan, J. King, M. Maclagan, and C. Watson, "The changing sound of the Māori language," in *An anthology on quantitative sociolinguistic studies of indigenous minority languages, Variationist Approaches to Indigenous Minority Languages*, J. Stanford and D. Preston, Eds. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2009, pp. 129–152.

[18] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and synthesising f0 contours with the discrete cosine transform," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 3973–3976.

[19] H. Fujisaki, "In search of models in speech communication research," in *Proceedings of Interspeech 2008*, 2008, pp. 1–10.

[20] J. King, M. Maclagan, R. Harlow, P. Keegan, and C. I. Watson, "The MAONZE project: changing uses of an indigenous language database," *Corpus Linguistics and Linguistic Theory 7(1)*, pp. 37–57, 2011.

[21] R. T. Sataloff and S. E. Linville, *Proffesional voice: The science and art of clinical care*. Plural Publishing, Inc., 2005, pp. 497–511.

[22] R. Colton, J. Casper, and R. Leonard, *Understanding voice problems: A physiological perspective for diagnosis and treatment*. Philadelphia, PA: Lippincott Williams & Wilkins, 2006.

[23] J. Harrington, *Phonetic analysis of speech corpora*. West Sussex, United Kingdom: Wiley-Blackwell, 2010.