A NEW FREQUENCY SHIFT FUNCTION FOR REDUCING INTER-SPEAKER VARIANCE

Christine Tuerk *Tony Robinson*
Cambridge University Engineering Department, Trumpington Street, Cambridge, England.

ABSTRACT

Speaker normalisation remains a very significant problem in speech research. One of the most immediate applications for a solution would be in the area of multiple-speaker speech recognition systems. These systems are faced with the task of assigning phonetic labels to portions of input speech, a task which is extremely complicated due to the enormous amount of variability within a phonetic class. Finding a good normalisation transformation would reduce this variability. Theoretical aspects of speech related studies would also benefit from a normalisation solution as it should lead to a greater understanding of the essential acoustic correlates that define a sound. A solution would aid researchers in the areas of perception and psycholinguistics. Normalisation techniques could also contribute to speech synthesis applications, especially in the area of producing multiple voices. This paper describes a frequency domain shift function which reduces the amount of inter-speaker variance within a phonetic class. The shift function is dependent upon the speaker's geometric mean pitch. The shift function is easily parameterised in a piece-wise linear fashion. Application of the shift allows a 15.8 - 17.0% reduction of variance. This reduction falls within 0.2% of the optimal pitch-only shift function for the data studied. In addition to variance reduction and recognition applications, this shift is easily applied as a means for warping speaker quality. This technique is applicable to synthesis systems where multiple voice qualities are desired.

1. INTRODUCTION

The scope of the speaker normalisation problem is vast. Men, women and children have physiologically very different vocal apparatus, and thus produce very different acoustic realisations of the same sound. Despite these different realisations, listeners are usually able to perceive and classify the sounds correctly. It is commonly accepted that in order to make this classification, the auditory system must make use of a transformation or series of transformations to the input signal in order to yield the correct phonetic assignment. Determining this exact transformation continues to elude researchers. This transformation seems to be dependent upon a number of interrelating factors including phonetic context, dialect, speaker pitch and acoustic stress. Ideally, a model unifying all these variables is needed. Before such a model can be found, it is instructive to study each of these factors individually. This paper focuses specifically on speaker pitch.

In this paper a new frequency shift function for reducing inter-speaker variance is explored. The next section will survey previous speaker normalisation work. The following sections will then present Miller's audio-perceptual theory [1] and examine how well his shift function reduces variance. In Section 5 a new shift function will be developed. Finally, Section 6 will present applications of the shift function and offer conclusions.

2. PREVIOUS WORK

An underlying question in speaker normalisation seems to be how speech sounds produced by varied vocal tracts can be perceived as "equivalent"? Some researchers are trying to answer this question by studying the perception of speech in the hope that understanding this problem will lead to insights into the type of auditory transformation humans perform on speech signals. These studies have also been used to support arguments regarding which portions of the speech signal are utilised in perception. The accepted viewpoint in the speech recognition community is that formants alone provide insufficient information for perception and that the entire short-term spectrum must be used in recognition. In [2], Bladon argues this case. Bladon's belief that formants alone are inadequate for perception has led him to work on a normalisation scheme involving a transformation of the entire spectrum [3]. His theory involves taking the entire short term spectra through a series of transformations to yield a pseudo-auditory spectra calibrated in sones/Bark versus Bark units. The resultant pseudo-spectra can then be used to normalise male and female speakers by a linear displacement on the Bark scale. However, the linear displacement required varied depending on the language and dialect under consideration [4].

Holmes [5] investigated the effect of the Bark scale frequency shift by synthesising data with various shifts ranging from 0.5 to 2.0 Barks. He found no obvious phonetic difference between the examples but claimed that the female voice quality was greatly inferior to that of the male. He felt that this inferiority may be due to either the inadequacy of a simple Bark shift to account for male and female differences, or that unaltered formant amplitudes could not represent correctly the results of a female glottal pulse.

Klatt [6] also performed a series of experiments in which parameters to synthesise speech were varied. Among the varied parameters were spectral tilt, relative formant amplitudes, high-pass, low-pass and notch-filtering, all of which turned out to have little phonetic relevance. He found that the most important variable in perception is the location of formant frequencies. These observations seem to indicate that one means of normalising between speakers is to find a method which reduces the variance in the location of their formant frequencies. This tack has been adopted by many researchers.

Many speaker normalisation efforts use the work of Peterson and Barney [7] as a starting point. Their work showed how the formant frequencies in the simple monophthongal vowels in a single context displayed considerable scatter. Attempts at reducing the scatter in vowel space has taken various forms. For example, Gerstman [8] normalised F1 and F2 frequencies by defining a range between the maximum and minimum F1 and F2 frequencies for each speaker. F1 and F2 frequencies were then linearly normalised based on these determined ranges. The normalised F1 and F2 values were then used in a vowel classification algorithm with improved results over unnormalised values. Another method for reducing the scatter between formant frequency values was developed by Wakita [9]. His method of normalising vowels used the length of the speaker's vocal tract. This length was determined automatically by solving linear prediction equations. Wakita's system is based on the assumption that for a given context, different speakers pronounce the same phoneme by having similar vocal tract configurations whose main difference is only in length. Thus, normalising the configurations to a reference length without altering the shapes should yield a set of very similar vocal-tract shapes for each vowel. This would then correspond to a smaller distribution of formant frequencies for a given vowel.

As Klatt's work has shown, formant frequencies are extremely important in perception, and yet, examining plots of F1 versus F2 (even after using normalisation techniques as described above) shows considerable overlap between vowels. The human perceptual system, as Bladon has suggested, must certainly be utilising more information than just formant frequency locations in performing the discrimination task. Many researchers have suggested that pitch is one of the additional factors that helps the listener to discriminate correctly between sounds. Potter and Steinberg [10] related formants to pitch. Fujisaki and Kawashima [11] have studied the role of pitch and higher formants on vowel perception and have found that perceptual normalisation is not complete unless both these factors are varied. Further evidence of pitch's importance is offered in the work of Traunmuller [12]. His research studied the perceptual relationship between one formant vowels and systematic variances of the pitch. His results showed that perception of the vowel quality varied with changes in pitch frequency.

3. THE AUDITORY-PERCEPTUAL THEORY

Another normalisation theory which combines the importance of formant locations and fundamental frequency with a theory on speech perception is found in the auditory-perceptual work of Miller. Simply stated, Miller's work implies that normalisation can be done by shifting frequency components by a factor proportional to the cube root of a speaker's geometric mean fundamental frequency (GMF0). His work descends from prior research with formant-ratio theories which state that the ratios between adjacent formant values (F2 to F1 and F3 to F2) can be used to indicate the identity of a vowel. These formant-ratio theories have been shown to reduce inter-speaker differences such as age and gender. However, one of the major weaknesses of theories based on formant-ratio methods is their inability to distinguish between some vowel pairs (for example, /aa/ and /ao/, and /uh/ and /uw/) because the vowel pairs produce very similar ratios.

In Miller's theory, input speech waveforms are transformed into short-term spectral analyses. From these short term spectra, four sensory pointers are derived. Three of these pointers, SF1, SF2, and SF3, correspond to the first three formant frequencies. A fourth pointer, called SR, is derived from the speaker's pitch and is defined as.

$$SR = GMF0_{ss} \sqrt[3]{\frac{GMF0_{cs}}{GMF0_{ss}}} \tag{1}$$

where $GMF0_{ss}$ is the geometric mean of a "standard" speaker's pitch and $GMF0_{cs}$ is the geometric mean of the current speaker's pitch. Miller defines the standard speaker's geometric mean pitch to be 168.0 Hz (on the basis that 168.0 is the geometric mean of the average adult male pitch, 125 Hz, and the average adult female pitch, 225 Hz). The four pointers are then used to calculate three new variables as follows:

$$y = \log SF1 - \log SR = \log \frac{SF1}{SR} \tag{2}$$

$$z = \log SF2 - \log SF1 = \log \frac{SF2}{SF1} \tag{3}$$

$$x = \log SF3 - \log SF2 = \log \frac{SF3}{SF2} \tag{4}$$

These three variables determine the perceptual response by yielding a set of coordinates in the auditory-perceptual space.

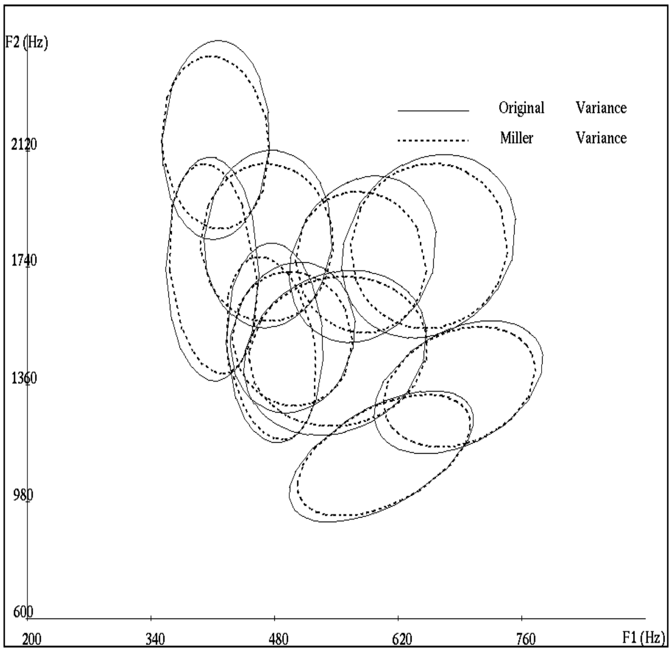

Figure 1. Variance ellipses by phoneme of original and Miller shifted data

Phonemic recognition occurs roughly by the "activation" of the target zones.

If the auditory-perceptual theory is correct, then for phonemic recognition to take place, differing instances of the same phoneme (produced by the same speaker or by different speakers) must produce spectral envelopes that yield values of x, y and z that fall into the perceptual target zone of that phoneme. This also has ramifications for normalisation – the spectral envelope can be shifted up and down in frequency, and, if the x, y, and z distances are maintained, the phonemic identity should remain the same. It must be noted that such a shift not only changes absolute formant locations, but should also change, via the sensory reference, the pitch of the utterance.

4. VARIANCE MEASUREMENT

The Miller shift was applied to the training data in the TIMIT [13] database. Each of the sentences was analysed for pitch contour and for formant locations. GMF0 and F1 and F2 values for each example of a monophthongal utterance were collected. The Miller shift function is based on equal numbers of adult male and female speakers giving a mean GMF0 of 168.0 Hz. However, the TIMIT database contains about three times as male as female speakers, giving a mean GMF0 of 141.3 Hz. To compensate, a classifier based on GMF0 was used as a male/female discriminator with a threshold of 168 Hz. The two speaker classes were weighted to give equal numbers of male and female speakers, so restoring the GMF0 to 168.4 Hz.

The variance measure was found through a full covariance analysis. Covariance analysis yields variance measurements (eigenvalues) along major and minor axes of the distribution. If v_{maj} and v_{min} denote the variances along the major and minor axes respectively, then an overall variance measure, V, which is proportional to the area of the distribution ellipse, can be given by

$$V = \sqrt{v_{min}} \times \sqrt{v_{mag}} \tag{5}$$

The Miller shift function was applied to the data set and the resultant variances were measured. Figure 1 shows the variance ellipses for each of the 10 monophthongal vowels for both the original data and for the Miller-shifted data. As can be seen, the Miller variances ellipses are smaller than their original data counterparts.

The Miller shift has created a slight change in mean F1 and F2 location for each vowel. This shift in mean locations affects the area of the ellipses and thus requires normalisation before a direct comparison with the original variances can be made. This shift is equivalent to a rotation followed by a scaling. The rotation causes no change in ellipse size but scaling does. This can be normalised by dividing by the determinant of the scaling matrix. The scaling matrix is

$$\begin{bmatrix} x_2/x_1 & 0 \\ 0 & y_2/y_1 \end{bmatrix} \tag{6}$$

where x_1 and y_1 denote the original mean locations and x_2 and y_2 denote the new mean. The determinant of the matrix is $x_2 y_2 / x_1 y_1$.

Figure 2 shows the normalised variance comparison across all phonemes of the original data and the data after the Miller shift was performed. The Miller shift clearly reduces the overall variance of the original data (from 21934 to 19207 a 12.4% drop), including a significant reduction (11.0%) in the standard deviation along the major axis. The overall standard deviation along the minor axis, however, only shows slight improvement with a 1.7% drop.

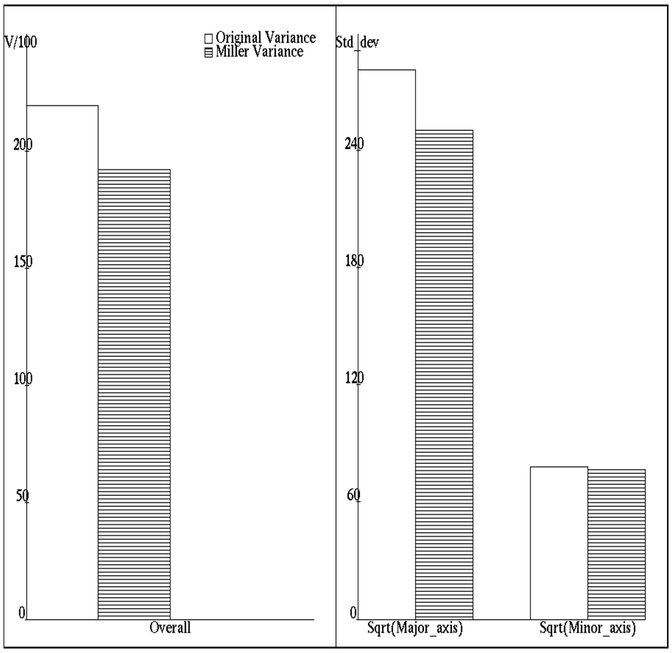

Figure 2. Variance of original data versus variance of Miller-shifted data

5. DERIVATION OF A NEW SHIFT FUNCTION

Miller based much of his work on the Peterson and Barney data. This database, however, consists of vowels spoken in only a single context and thus is a lot smaller than the TIMIT database. Studying the characteristics of the larger TIMIT database may yield a shift function which gives better results than the Miller shift.

Figure 3 displays a graph with data points corresponding to the optimal shift factor for reducing the distance between the measured F1 and F2 values in comparison with the average F1 and F2 values. The data points are weighted averages obtained over all 10 monophthongal vowels. The two straight lines indicate the best-fit regression lines for two regions. The curved line displays the Miller shift function. The regression line function appears to fit the data much better than the Miller function, especially in the region of low-pitched speakers. The shift function can be summarised by the following 2 equations:

$$\begin{array}{lll} y = -0.000891x + 1.221 & \text{if} & x <= 168 \\ y = -0.000579x + 1.083 & \text{if} & x > 168 \end{array} \tag{7}$$

where x is the geometric mean of the pitch and y is the shift factor amount.

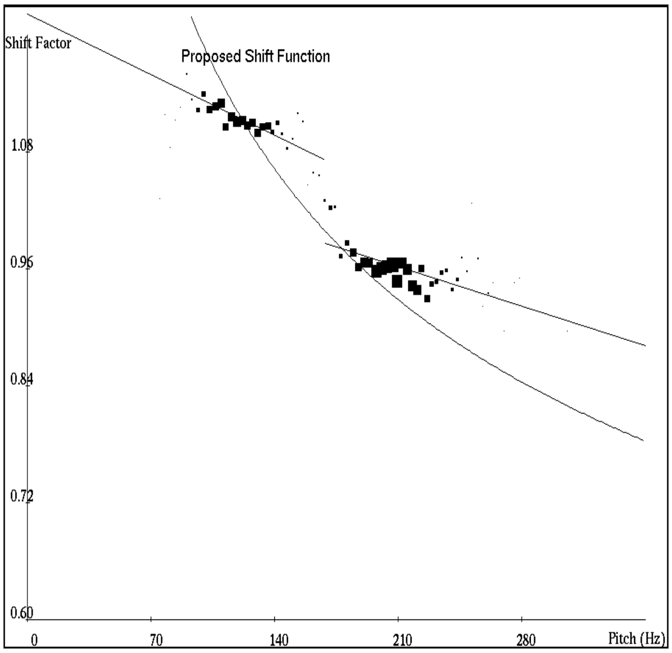

Figure 3. Best-fit shift factor functions

Applying this shift function to the TIMIT data yields a greater decrease in the overall variance than that given by the Miller shift. Figure 4 shows a bar graph of the variances including the original, Miller shifted, new and optimal variances. The optimal variance is defined as the lowest possible variance that can be obtained using only F1, F2 and pitch as variables. The optimal shift function is found by storing the shift factor for each integral pitch value (each of the dots in Figure 3) in a table. The new shift, with only 5 parameters, comes very close to matching the optimal shift. V for the new shift yields 18213.08, or a 17.0% improvement over the original data (the Miller shift yielded a 12.4% improvement). The optimal shift would yield a V measurement of 18158.04, or a 17.2% improvement. Not only does the new shift give a bigger reduction in overall variance than the Miller shift, it also shows a significant reduction in the standard deviation of both the major and minor axes.

The new shift function was also applied to the unseen TIMIT test data. Again, the new shift function performed better than the Miller shift function. Figure 5 shows the results in bar graph form. The Miller shift function performed similarly on the training and test sets, yielding an 11.8% variance decrease overall. Again, the Miller shift displayed very little improvement in the standard deviation along the minor axis (only 1.3% improvement). The new shift function again gave better results than the Miller function with a 15.8% decrease in overall variance. Again, the new shift function was also able to reduce the standard deviation along the minor axis better than the Miller shift (a 5.0% reduction).

6. APPLICATIONS AND CONCLUSIONS

The shift function determined above can be applied as a means to reduce variance. This reduction in variance should aid recognition. In order to illustrate this, a simple Gaussian classifier was built and used as a recogniser on the training and test data sets. (It is not claimed that a Gaussian classifier is optimal, merely that its use may demonstrate how recognition rate may be improved through the application of the shift function). The classifier built on the original training data resulted in a recognition rate of 45.8% on the training data and 44.1% on the test data. Building a classifier on the shifted training data, and using it to classify the shifted training and test sets,

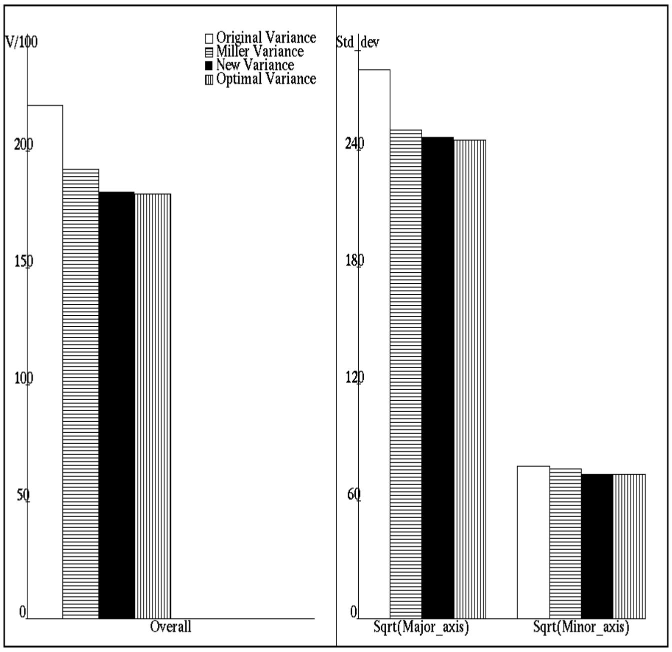

Figure 4. Variances of original data, Miller shift, new shift and optimal shift

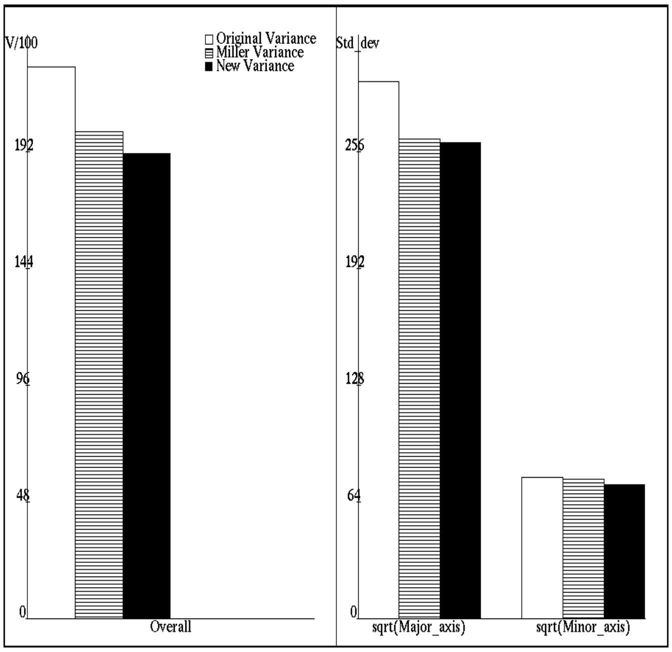

Figure 5. Variances of original, Miller and new shift data - test set

yielded results of 48.4% and 47.1% respectively. This approximate 3% recognition improvement was slightly better than that obtained on a classifier built on the Miller shifted data.

The shift function may also be used in speech synthesis applications. Here, it can be used to transform sentences to take on varied speaker qualities. That is, the formants and pitch of a sentence spoken by a female speaker can be shifted down to the formants and pitch more characteristic of a male speaker. Similarly, the speech of a low-pitched male can be transformed to take on female speaker quality. This technique can be used to change the character of sentences in the multiple-speaker TIMIT database to a more "generic" speaker for parameter gathering purposes; conversely, the inverse of the shift function can be used to change the output of a single-speaker synthesiser to the qualities of different speakers.

This paper has proposed a shift function which reduces the variance in formant frequency location. This function relies on pitch and performs better than the Miller function, another function which uses pitch as the independent variable. The shift function shows improvement in a simple recognition task. The shift is also useful for changing voice quality. This has applications for speech synthesis where the production of multiple voice qualities normally assumes duplication of effort in synthesis database gathering.

It is fully realised that pitch is not the only variable which effects the location of formant frequencies. Context, stress, dialect, etc. are also factors. Ideally, a model which integrates all these variables into a coherent model should be explored.

7. ACKNOWLEDGEMENTS

This work has been supported in part by the National Science Foundation and the UK Science and Engineering Research Council.

REFERENCES

[1] J.D. Miller. Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, 85:2114–2134, 1989.

[2] A. Bladon. Arguments against formants in the auditory representation of speech. In R. Carlson and B. Granstrom, editors, *The Representation of Speech in the Periphereal Auditory System*. Elsevier Biomedical Press, 1982.

[3] R.A.W. Bladon. Problems of normalizing the spectral effects of variations in the fundamental. In *Proceedings of the Institute of Acoustics Autumn Conference*, pages A5.1–A5.5, 1982.

[4] A. Bladon. Acoustic phonetics, auditory phonetics, speaker sex and speech recognition: A thread. In F. Fallside and W. Woods, editors, *Computer Speech Processing*. Prentice Hall, 1985.

[5] J.N. Holmes. Normalization in vowel perception. In J.S. Perkell and D.H. Klatt, editors, *Invariance and Variability in Speech Processes*. Lawrence Erlbaum Associates, Publishers, 1986.

[6] D.H. Klatt. Prediction of perceived phonetic distance from critical-band spectra: A first step. In *Proceedings of the Int. Conf. Acoust. Speech Signal Processing*, pages 1278–1281, 1982.

[7] G.E. Peterson and H.L. Barney. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24:175–184, 1952.

[8] L.J. Gerstman. Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, AU-16:78–80, 1968.

[9] H. Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-25:183–192, 1977.

[10] R.K. Potter and J.C. Steinberg. Toward the specification of speech. *Journal of the Acoustical Society of America*, 22:807–820, 1950.

[11] H. Fujisaki and T. Kawashima. The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, AU-16:73–77, 1968.

[12] H. Traunmuller. Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, 69:1465–1475, 1981.

[13] L.F. Lamel, R.H. Kasel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 26–32, 1987.