Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Low-dimensional, auditory feature vectors that improve vocal-tract-length normalization in automatic speech recognition

Jessica Monaghan, Christian Feldbauer, Tom Walters and Roy Patterson

Centre for the Neural Basis of Hearing, Department of Physiology, Development and
Neuroscience, University of Cambridge, Downing Site, CB23EG Cambridge, UK
rdp1@cam.ac.uk

The syllables of speech contain information about the vocal tract length (VTL) of the speaker as well as the glottal pulse rate (GPR) and the syllable type. Ideally, the pre-processor for automatic speech recognition (ASR) should segregate syllable-type information from VTL and GPR information. The auditory system appears to perform this segregation, and this may be why human speech recognition (HSR) is so much more robust than ASR. This paper compares the robustness of recognizers based on two types of feature vectors: mel-frequency cepstral coefficients (MFCCs), the traditional feature vectors of ASR, and a new form of feature vector inspired by the neural patterns produced by speech sounds in the auditory system. The speech stimuli were syllables scaled to have a wide range of values of VTL and GPR. For both recognizers, training took place with stimuli from a small central range of scaled values. Average performance for MFCC-based recognition over the full range of scaled syllables was just 73.5%, with performance falling to 4% for syllables with extreme VTL values. The bio-acoustically motivated feature vectors led to much better performance; the average for the full range of scaled syllables was 90.7%, and performance never fell below 65%.

1 Introduction

When an adult and a child say the same sentence, the information content is the same, but the waveforms are very different. Adults have longer vocal tracts and heavier vocal cords than children. Despite these differences, humans have no trouble understanding speakers with varying vocal tract lengths (VTLs) and glottal pulse rates (GPRs); indeed, [1] showed that both VTL and GPR could be extended far beyond the ranges found in the normal population without a serious reduction in recognition performance. This robustness of human speech recognition (HSR) stands in marked contrast to that of automatic speech recognition (ASR), where recognizers trained on an adult male do not work for women, let alone children [2].

GPR and VTL are properties of the source of the sound at the syllable level in speech communication, quite separate from the information that determines syllable type. The microstructure of the speech waveform reveals a stream of glottal pulses each followed by a complex resonance showing the composite action of the vocal tract above the larynx on the pulses as they pass through it. The resonances of the vocal tract are known as formants and they determine the envelope of the short-term magnitude spectrum of speech sounds. The formant peak frequencies are determined partly by vocal tract shape and partly by VTL, which is strongly correlated to height in humans [3]. As a child grows into an adult, the formants of a given vowel decrease in inverse proportion to VTL, and this is the form of VTL information in the magnitude spectrum [4]. When plotted on a logarithmic frequency scale the frequency dilation produced by a change in speaker size becomes a linear shift of the spectrum, as a unit, along the axis – towards the origin as the speaker increases in size. This paper is concerned with the robustness of ASR when presented with speakers of widely varying sizes; that is, how the performance of an ASR system varies as the spectra of speech sounds expand or compress in frequency with changes in speaker size.

ASR requires a compact representation of speech sounds for both the training and recognition stages of processing, and traditionally ASR systems use a frame-based spectrographic representation of speech to provide a sequence of ‘feature vectors’. Ideally the construction of the feature vectors should involve segregating the syllable type information from the speaker-size information (GPR and VTL), and the removal of the size information from the feature vectors. In theory, this would help make the recognizer robust to variation in speaker size.

1.1 Encoding of VTL information in MFCC feature vectors

Most commercial ASR systems use mel-frequency cepstral coefficients (MFCCs) as their feature vectors because they are believed to represent speech information well, and they are robust to background noise. A mel-frequency cepstral coefficient is the amplitude of a cosine function fitted to a spectral frame of a sound (plotted in quasi-log-frequency, log-magnitude coordinates). The MFCC feature vector is computed in a sequence of steps. (1) A temporal window is applied to the sound and a fast Fourier transform is performed on this windowed signal. (Note that the window position is stepped regularly along in time without regard to the timing of the glottal pulses.) (2) The spectrum is mapped onto the mel-frequency scale using a triangular filter-bank; the mel-frequency scale is a quasi-logarithmic scale, similar to Fletcher’s ‘critical band’ scale or the ERB scale [5]. (3) Spectral magnitude is converted to the logarithm of spectral magnitude. (4) A discrete cosine transform (DCT) is applied to the mel-frequency log-magnitude spectrum. The MFCCs are the coefficients of this cosine series expansion, or ‘cepstrum’. (5) The first twelve of these cepstral coefficients form the feature vector; the remaining higher-order coefficients are discarded, which has the effect of smoothing the mel-frequency spectrum as the feature vector is constructed. A zeroth coefficient is appended, proportional to the log energy.

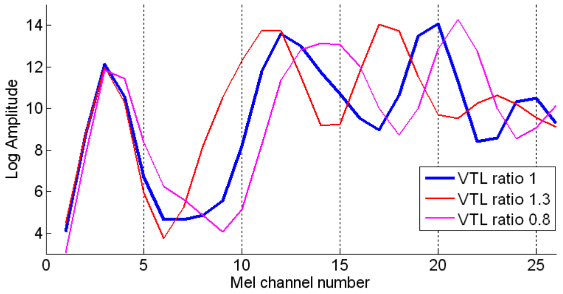

Fig. 1 The smoothed spectra of three scaled versions of the same vowel produced using the 26-channel mel-frequency ‘filterbank’. The spectrum shifts upwards in frequency as the VTL ratio reduces from 1.3 (red) to 0.8 (magenta). This shift is approximately linear in the higher channels, but not for low channels.

Despite their popularity, cepstral coefficients generated with a discrete cosine transform have an intrinsic flaw when used to represent speech sounds, which are known to vary in acoustic scale. As illustrated in Fig. 1, for a given vowel, a change in acoustic scale (or VTL) essentially results in a shift of the spectrum on the mel-frequency axis. Since MFCCs are generated with a cosine transform, the basis functions are specifically prohibited from shifting with acoustic scale. That is, a cosine maximum which fits a formant peak for a vowel from a given vocal tract with a specific length, cannot shift to follow the formant peak as it shifts with VTL; the maxima for a given cosine component occur at multiples of one specific frequency and they cannot be shifted. As a result, a change in speaker size leads to a change in the magnitude of all of the cepstral coefficients, whereas, a change in the phase of the basis functions would provide a more consistent representation of syllable type. The size information is still present in the MFCCs but it is not easily accessible, and as a result, a very large database and excessively long training times would be needed to accurately model the sound.

Most existing techniques for VTL normalization of MFCC feature vectors involve attempts to counteract the dilation of the magnitude spectrum by warping the mel 'filters' (the weighting functions that convert the magnitude spectrum of the vowel into a mel-frequency spectrum) prior to the production of the cepstral coefficients. Unfortunately, the process of finding the value of the relative size parameter is computationally expensive, and must be done individually for each new speaker. The problem is that the relationship between the value of a specific MFCC and VTL is complicated, and a change in VTL typically leads to substantial changes in the values of all of the MFCCs. In other words, the individual cepstral coefficients all contain a mixture of syllable-type information and VTL information, and it is very difficult to segregate the information once it is mixed in this way. As a result, the MFCCs do not themselves effect segregation of the two types of information. The segregation and normalization problems are left to the recognizer that operates on the feature vectors, and it is this which limits the robustness of ASR to changes in speaker size when it is based on MFCC feature vectors [6].

1.2 AIM feature vectors

The Auditory Image Model (AIM) [7] simulates the general auditory processing that is applied to speech sounds, like any other sounds, as they proceed up the auditory pathway to the speech specific processing centers in the temporal lobe of cerebral cortex. AIM produces a pitch-invariant, size-covariant representation of sounds referred to as the size-shape image (SSI). This representation includes a simulation of the normalization for acoustic scale that is assumed to take place in the perception of sound by humans. The SSI is a 2-D representation of sound with dimensions of 'auditory filter frequency on a quasi-logarithmic (ERB) axis' by 'time-interval within the glottal cycle.' The SSI can be summarized by its spectral profile [8], and the profile has the same scale-shift covariance properties as the SSI itself.

The SSI profiles produced by the three /i/ vowels described above are shown in Fig. 2. They are like excitation patterns [5], or auditory spectra, and the figure shows that the distribution associated with the vowel /i/ shifts along the axis with acoustic scale. Thus, the transformations performed by the auditory system produce segregation of the complementary features of speech sounds; that is, the information about the size of the speaker, and the size invariant properties of speech sounds, like vowel type. In this way the transformations simulate the neural processing of size information in speech by humans. Experiments show that speaker size discrimination and vowel recognition performance are related: when discrimination performance is good, vowel recognition performance is good [1]. This suggests that recognition and size estimation take place simultaneously. It is assumed that the acoustic scale information is essentially VTL information, and that it is used to evaluate speaker-size, and that the normalized shape information facilitates speech recognition and makes the recognition processing robust.

Section 2.3 shows how the information content of the SSI profile can be summarized with a mixture of four Gaussians to produce a four dimensional feature vector. The performance of a recognizer using these bio-acoustically motivated feature vectors is compared with that of a recognizer using traditional MFCCs to demonstrate the greater robustness of feature vectors based on auditory principles.

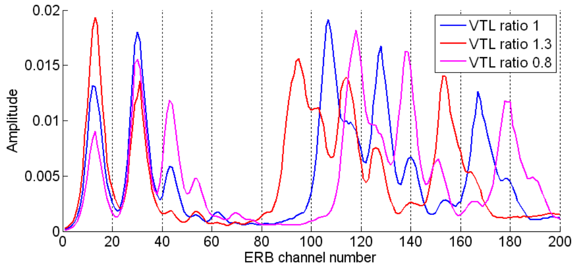

Fig. 2 SSI profiles of three scaled versions of the vowel /i/ with a GPR of 165Hz for a 200 channel ERB filter-bank. Apart from frequency effects in the lower channels there is a clear linear shift of the vowel spectrum with VTL ratio.

2 METHOD

The speech corpus used in this study was compiled by Ives et al. (2005) [9] who used phrases of four syllables to investigate VTL discrimination. There were 180 syllables in total, composed of 90 consonant-vowel and vowel-consonant pairs.

The syllables were recorded from one speaker (author RP) in a quiet room with a Shure SM58-LCE microphone. The microphone was held approximately 5 cm from the lips to ensure a high signal to noise ratio and to minimize the effect of reverberation. A high-quality PC sound card (Sound Blaster Audigy II, Creative Labs) was used with 16-bit quantization and a sampling frequency of 48 kHz. The syllables were normalized by setting the RMS value in the region of the vowel to a common value so that they were all perceived to have about the same loudness.

2.1 Scaling the syllable corpus

Once the syllable recordings were edited and standardized, a vocoder referred to as STRAIGHT [10] was used to generate all the different 'speakers,' that is, versions of the corpus in which each syllable was transformed to have 57 combinations of VTL and GPR. The central speaker was assigned a GPR of 171.7 Hz and a VTL of 146.9 mm, which was chosen to be midway on the line between the average logGPR-logVTL values for men and women. For scaling purposes, the VTL of the original speaker was taken to be 165 mm. The average values of VTL were taken from [3] and the average GPR was taken from [11].

A set of 56 scaled speakers were produced with STRAIGHT in the region of the GPR-VTL plane surrounding the central speaker, and each speaker had one of the combinations of GPR and VTL illustrated by the points on the radial lines of the GPR-VTL plane in Fig. 3. There were seven speakers on each of eight spokes. The ends of the radial lines form an ellipse whose minor radius is four semi-tones in the GPR direction and whose major radius is six semi-tones in the VTL dimension. The seven speakers along each spoke are spaced logarithmically in this log-log, GPR-VTL plane. The spoke pattern was rotated anti-clockwise by 12.4 degrees so that there was always variation in both GPR and VTL when the speaker changed. This angle was chosen so that two of the spokes form a line coincident with the line that joins the average man with the average woman in the GPR-VTL plane.

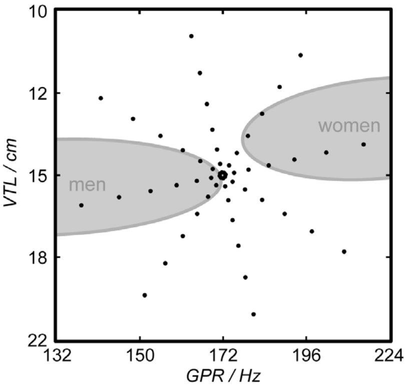

Fig. 3 The locations of the scaled speakers in the GPR-VTL plane: The GPR of the scaled speaker varied between 137 and 215 Hz; the VTL varied between 11 and 21 cm. The central speaker had a GPR of 172 Hz and a VTL of 15 cm. The grey ellipses correspond to speakers in the normal population as modelled by [12].

2.2 The Hidden-Markov Model Toolkit

The hidden Markov model tool kit (HTK) [13] was used as a platform to produce the recognizers. HTK models speech as a sequence of stationary segments, or frames, produced by a hidden Markov model (HMM). For an isolated syllable recognizer, one HMM is used to model the production of each syllable. In all of the experiments in this paper, the HMM recognizers were trained on the reference speaker of the scaled-syllable database, and the eight speakers closest to the reference speaker in the GPR-VTL plane. This procedure was intended to imitate the training of a standard, speaker-specific ASR system, which is trained on a number of utterances from a single speaker. The eight adjacent points provided the small degree of variability needed to produce a stable model for each syllable. The recognizers were then tested on all of the scaled speakers, excluding those used in training, to provide an indication of their relative performance.

The audio files used for training were converted to HTK files consisting of frames of either MFCCs or AIM feature vectors. These HTK files were labeled by syllable and the parameters of each syllable model, such as the output distribution and the transition probability of each state, were estimated from the nine HTK files in the training set. In the testing stage the most probable HMM that produced each file in the rest of the corpus was found, and the file was assigned the syllable corresponding to that HMM as its transcription. The transcriptions generated were then compared to the true transcription or ‘labels’ of the files and a recognition score was calculated.

An HMM with three emitting states was used for both recognizers; three emitting states is sufficient for single syllables. The HMM topology was varied and the optimal recognition values were found for both recognizers

2.3 Summarizing the formant frequency information of the profiles in a low-dimensional feature vector

SSI profiles, as described in section 1.2, were produced for 10-ms frames of each syllable file in the scaled syllable corpus. The profiles were produced using AIM-C, an implementation of AIM in C++. Feature vectors were produced by first applying power-law compression with an exponent of 0.8 to the profile magnitude and normalizing them to sum to unity. The profiles were treated like probability density functions and a modified expectation-maximization (EM) algorithm was used to fit a mixture of four Gaussians to the profiles. The parameters of this mixture of Gaussians make up the components of the low-dimensional feature vectors. The motivation for this technique can be understood by looking at the fit to the vowel /i/ shown in Fig. 4. There are three main concentrations of energy in vowels and sonorant consonants and they have a roughly Gaussian shape. These are encoded by three of the Gaussians, while the remaining Gaussian encodes a gap in the spectrum between the first and second formants.

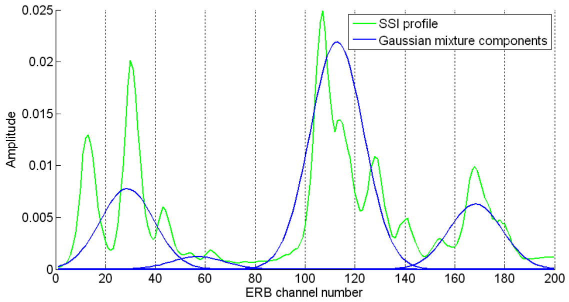

Fig. 4 Illustration of the feature extraction process: Four Gaussians (blue) with fixed variances were fitted to the SSI profile (green) of an /i/ vowel using an EM algorithm and a minimum-separation limitation. The feature vector is constructed from three of the four Gaussian weights plus a log-energy term.

To get a more consistent fit, the EM algorithm is modified in three ways: (1) the variance of the Gaussian is not updated but remains at the original value of 115 channels squared (2) the conditional probabilities of the mixture components in each filterbank channel are expanded according to a power-law (with an exponent of 0.6) and re-normalized in each iteration to reduce the overlap between Gaussians, and (3) an initialization step is introduced. Having a fixed variance reduces the number of degrees of freedom, resulting in a more consistent fit. The optimal value of the variance was established during preliminary experiments using only the vowels. Having wide Gaussians was found to prevent the fitting of Gaussians to individual resolved harmonics, reducing sensitivity to pitch variation. The initialization step fits two Gaussians to the profile, and uses the interval between the means of these Gaussians to provide an initial position for the four Gaussians in the second stage. The features themselves were the weights of the four Gaussians, which since they sum to one can be summarized as three parameters. The log of the energy of the un-normalized profile was included in the feature vector. Recognition performance over the vowels was then 100%. It is this method that was used to produce the recognition results with the auditory pre-processor reported in the next section.

In summary, a four-dimensional, auditory feature vector was produced using the logarithm of the energy of the original profile, and three of the Gaussian weights. First and second difference coefficients were computed between temporally adjacent feature vectors and added to the feature vector in all cases. Thus, the length of the AIM feature vectors passed to the recognizer was 12 components, whereas it was 39 components for the MFCC feature vectors. Having feature vectors with a lower dimensionality should reduce the time taken to run the training and recognition algorithms substantially in full scale systems.

3 Results and Discussion

3.1 HMM recognizer operating on MFCC feature vectors

In the initial experiment with the MFCC feature vectors, the recognizer was based on an HMM with a topology that had three emitting states and a single Gaussian output distribution for each state. The recognizer was trained on the original speaker and the eight speakers on the smallest ellipse nearest to the original speaker. The average recognition accuracy for this configuration, over the entire GPR-VTL plane, was only 65.0 %. To ensure that the results were representative of HMM performance, a number of different topologies were trained and tested. Performance was best for an HMM topology consisting of four emitting states, with several Gaussian mixtures making up the output distributions for each emitting state. The number of training stages was also varied to avoid over-training. The optimum performance, using the best topology, was 73.5 % after nine iterations of the training algorithm. A further experiment was carried out using MFCCs produced from a 200 channel mel filterbank to check that the performance of the recognizer was not being limited by a lack of spectral resolution. The performance using these MFCCs was 67.7 % for the initial topology with three emitting states and 73.3 % using the best topology from the previous experiments, indicating that 26-channel resolution was not a serious limitation.

The performance for all of the individual speakers, using this topology, is shown in Fig. 5. There is a central region adjacent to the training data for which performance is 100 %; it includes the second ellipse of speakers and several speakers along spokes 1 and 5 where VTL does not vary much from that of the reference speaker. As VTL varies further from the training values, performance degrades rapidly. This is particularly apparent in spokes three and seven, where recognition falls close to 0 % for the extremes, and to a lesser extent on spokes two, four, six and eight. This demonstrates that this MFCC recognizer cannot extrapolate beyond its training data to speakers with different VTLs. In contrast, performance remains consistently high along spokes 1 and 5, where the main variation is in GPR. This is not surprising since the process of extracting MFCCs eliminates most of the GPR information from the features. This figure shows the performance that sets the standard for comparison with the auditory feature vectors.

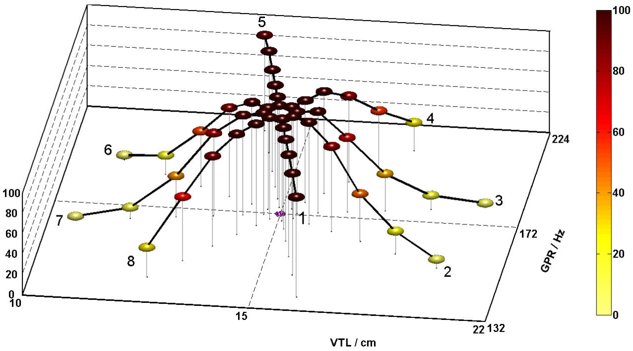

Fig. 5 Performance of the MFCC recognizer for individual speakers across the VTL-GPR plane. The training set was the reference speaker and the eight surrounding speakers of the smallest ellipse. Performance is seen to deteriorate rapidly as VTL diverges from that of the training region. Average performance using this optimum topology was 73.5 %.

3.2 HMM recognizer operating on AIM feature vectors

In the initial experiment with the AIM feature vectors, the recognizer was based on an HMM with a topology that had three emitting states and a single Gaussian output distribution for each state, as for the MFCC recognizer. The initial recognition rate using the SSI feature vectors was 84.6 % over the full range of speakers across the GPR-VTL plane; this is well above the initial performance with MFCC feature vectors. Performance was best for an HMM topology consisting of two emitting states, with several Gaussian mixtures making up the output distributions for each emitting state. The number of training stages was again varied. After optimization of the topology and nine iterations of the training algorithm, performance rose to 90.7 %, which is well above the 73.5 % achieved after similar optimization with the MFCC feature vectors.

Performance obtained using this topology for the individual speakers across the GPR-VTL plane, is shown in Fig. 6. As with the MFCC recognizer, performance is best along spokes one and five. However, unlike the MFCC recognizer, performance along most of the spokes is near ceiling after optimization. The worst performance, for the speaker at the end of spoke three, was 66.5 %, which compares with 3.8 % in the MFCC case. There is a drop in performance at the extremes of spokes three and seven, although the drop is small in comparison to that seen in the

MFCC case. The results indicate that there is still some sensitivity to change in VTL in the AIM feature vectors. Since it affects only the extreme VTL conditions, it seems likely that it is due to edge effects at the Gaussian fitting stage. That is, when a formant occurs near the edge of the spectrum, the tail of the Gaussian used to fit the formant prevents it from shifting sufficiently to center the Gaussian on the formant. If this proves to be the reason, it suggests that performance is not limited by the underlying auditory representation (the SSI) but rather by a limitation in the feature extraction process – a limitation that should be amenable to improvement.

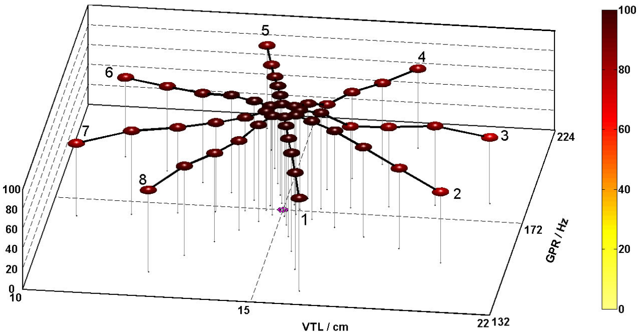

Fig.6 Performance of the AIM recognizer for individual speakers across the VTL-GPR plane. The training set was the same as in the MFCC case. Performance only deteriorates for speakers with extreme VTL values. Average performance using this optimum topology was 90.7 %.

4 Conclusion

In an effort to improve the robustness of ASR recognizers to variation in speaker size, a new form of feature vector was developed, based on the spectral profiles of the SSI stage of the auditory image model (AIM). The value of the new feature vectors was demonstrated using an HMM syllable recognizer, which was trained on a small number of speakers with similar GPRs and VTLs, and then tested on speakers with widely different GPRs and VTLs. Performance was compared to that of a traditional ASR system operating on MFCC feature vectors. When tested on the full range of scaled speakers, performance with the AIM feature vectors was shown to be significantly better (~91 %) than that with the MFCC feature vectors (~74 %). Moreover, the auditory feature vectors are far smaller (12 components) than the MFCC feature vectors (39 components). The study demonstrates that the high resolution, spectral profiles typical of auditory models can be successfully summarized in low-dimensional feature vectors for use with recognition systems based on standard HMM techniques.

Acknowledgments

The research was supported by the UK Medical Research Council [G0500221, G9900369] and the European Office of Aerospace Research & Development (EOARD) [FA8655-05-1-3043]. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of EOARD or the MRC.

References

[1] D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, T. Irino, "The processing and perception of size information in speech sounds", *J. Acoust. Soc. Am.* 117, 305-318 (2005).

[2] A. Potamianos, S. Narayanan, S. Lee, "Automatic speech recognition for children". *Proceedings of European Conference on Speech, Communication and Technology*, Rhodes, Greece, pp. 2371–2734 (1997).

[3] W.T. Fitch, J. Giedd, " Morphology and development of the human vocal tract: a study using magnetic resonance imaging", *J. Acoust. Soc. Am.* 106, 1511-1522 (1999).

[4] R. D. Patterson, D. R. R. Smith, R. van Dinther, T. C. Walters, "Size Information in the Production and Perception of Communication Sounds". In W. A. Yost, A. N. Popper and R. R. Fay (Eds.), Auditory Perception of Sound Sources. Springer US, 43-75 (2006)

[5] B. R. Glasberg, B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data", *Hearing Research* 47, 103-138 (1990).

[6] C. Feldbauer, J. J. M. Monaghan, R. D. Patterson, "Continuous Estimation of VTL from Vowels Using a Linearly VTL-Covariant Speech Feature", Acoustics'08. Paris (2008)

[7] T. Irino, R. D. Patterson, "Segregating Information about Size and Shape of the Vocal Tract using the Stabilised Wavelet-Mellin Transform", S*peech Communication* 36, 181-203 (2002).

[8] R. D. Patterson, R. van Dinther, T. Irino, "The robustness of bio-acoustic communication and the role of normalization", P*roc. 19th International Congress on Acoustics*, Madrid, Sept, ppa-07-011 (2007).

[9] D. T. Ives, D. R. R. Smith, R. D. Patterson, "Discrimination of speaker size from syllable phrases", *J. Acoust. Soc. Am.* 118 (6), 3816-3822 (2005).

[10] H. Kawahara, T. Irino, "Underlying principles of a high-quality, speech manipulation system STRAIGHT, and its application to speech segregation". In P. Divenyi (Ed.), Speech separation by humans and machines. Kluwer Academic: Massachusetts, 167-179 (2005).

[11] G.E. Peterson, H.L Barney, "Control methods used in a study of the vowels", *J. Acoust. Soc. Am.* 24, 175-184 (1952).

[12] R. E. Turner, R.D. Patterson, "An analysis of the size information in classical formant data: Peterson and Barney (1952) revisited", *The Acoustical Society of Japan* 33, No. 9, 585-589 (2003).

[13] S. Young, G.Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, "The HTK Book (for HTK version 3.4)", Cambridge University Engineering Department, Cambridge (2006).