A NEW SPEECH ANALYSIS SYSTEM: ASSESS (AUTOMATIC STATISTICAL SUMMARY OF ELEMENTARY SPEECH STRUCTURES)

R. Cowie, M. Sawey and E. Douglas-Cowie
School of Psychology / School of English, Queen's University, Belfast, UK

ABSTRACT

ASSESS is a set of programs which automatically segment the speech signal, and then provide systematic statistical measures of the structures elicited. The attributes that it describes are relevant to distinguishing a number of speech varieties, normal and abnormal. We have applied it to speech pre and post cochlear implant, schizophrenic speech, and markers of emotion in speech.

INTRODUCTION

This paper is about extending a natural approach to measuring speech properties. The general aim is to develop systems which analyse speech automatically and create a battery of statistical descriptors. The distinctive extension that we consider is to incorporate preprocessing which extracts from the raw signal key features and relationships, whose properties can then be measured and summarised. This has the potential to produce much richer descriptions than approaches which treat the signal as a whole. We have developed a system which uses this strategy, and tested its value in several projects. It is called ASSESS, as an acronym for Automatic Statistical Summary of Elementary Speech Structures.

ASSESS grew out of concern to measure speech attributes which lead listeners to place people in emotionally and socially significant categories. It is not self-evident that statistical measures are relevant to the problem: they might be dominated by idiosyncratic features of speech at one extreme, or by phonetic and syntactic structure at the other. However, a good deal of evidence suggest otherwise. ASSESS builds on that evidence, and incorporates relevant measures in a systematic framework.

Its immediate origins lie in research on the way speech changes when people become deaf. We were able to show that listeners made distinctive (and usually adverse) judgements about deafened people's speech [1], but our phonetic analysis, using a traditional segmental analysis, showed so few abnormalities that our data were used to argue that there were no significant abnormalities [2].

We turned to statistical techniques because impressionistically, spectrograms of deafened and control speakers looked quite grossly different. Studies confirmed that such an approach captures some effects of acquired deafness on speech [3], [4], particularly the way intensity rises and falls. Patterns of change in F0 also appear to distinguish deafened and hearing speakers statistically: the distribution of types is more obviously anomalous than the nature of individual tokens [5]. A less intuitive feature of deafened speech involves spectral balance. Unusually high proportions of energy concentrate around the region associated with F2 [3].

There are indications that other speech varieties may be distinguished in related ways. Formal and informal speech styles appear to differ statistically [3]. In Belfast speech at least, social distinctions involve the distribution of pitch patterns which invite summary in statistical terms [6]. Certain forms of schizophrenia seem to be marked by the statistical distributions associated with F0 [7]. Statistical properties related to intonation bear at least broad relationships to the vocal expression of emotion [8],[9],[10]. Voice quality has often been identified as a factor in reactions to and evaluations of a speaker [11], and some types of voice involve distinctive balances between broad regions of the spectrum [12].

ASSESS develops the intuition that these observations reflect a domain which is both coherent and significant. It is a prototype which is competent enough to let us probe the nature of the domain and its relation to the way people respond to voices. Its structure offers a useful overview of the tasks that systems of this kind need to address. Its performance has allowed us to carry out reasonably large scale studies, which help to validate the approach, and to refine it in the light of experience.

SYSTEM OVERVIEW

ASSESS has four components.
(1) InASSESS saves samples from tape using a CED 1401 signal capture unit. Sampling is at 20kHz: frequencies above 10kHz are removed by low pass analogue prefiltering. Memory limitations restrict a single sample to 8.2 seconds. As typical inputs are much longer, they are divided by hand using a display of the time waveform to find natural cut-off points.
(2) QuantASSESS is a signal processing stage which extracts basic descriptions - an intensity profile, initial estimates of F0, and a 1/3 octave spectrum.
(3) QualASSESS takes the output of QuantASSESS, identifies key features, and creates initial statistical descriptions.
(4) SumASSESS integrates data from files which form a single passage and generates graphical and numerical summaries. Additional programs integrate data from multiple passages and replay stored graphical representations.
QuantASSESS and QualASSESS are currently written in TurboPascal and run on IBM PCs. Both can run a large batch of files from InASSESS, though each one is analysed independently. SumASSESS and supporting programs are written in QuickBasic and run on a Macintosh (allowing access to Macintosh statistical packages and graphics: convenience at that level is important).

Hardware seriously limits the utility of the current system. Capture is the greatest bottleneck, and work is in hand on a replacement using a SUN SparcII to capture long samples directly. The other components will then be transferred to the same platform and adjusted accordingly.

EXTRACTING KEY FEATURES

QuantASSESS has three functions.
(1) It recovers intensity from voltage. Energy is integrated over periods of 25.6 milliseconds, which we call 'slices'.
(2) It creates a spectrum. An FFT is calculated for each slice, and used to estimate the output of 18 1/3 octave filters, with centre frequencies running from 128Hz to 6.5kHz.
(3) It makes initial estimates of voice pitch using an algorithm which looks for upstrokes in the time waveform that may signal vocal cord openings. The intervals between these strokes give direct but noisy pitch estimates. QuantASSESS then finds sequences of strokes which could correspond to evenly spaced vocal cord openings, allowing that a few real strokes may have been missed, or spurious strokes introduced. Confidence values are associated with pitch estimates derived from these sequences: they depend on the number of missing or spurious strokes assumed.

QualASSESS recovers descriptions of key contours, and finds key transition points - beginnings and ends of silences and fricative bursts, maxima and minima on the intensity and F0 contours, and plateau boundaries. The idea of plateaux derives from work on schizophrenic speech [13]. They are 'flat' regions around an inflection where contour height has changed by less than 5% of the distance to the next inflection. Falls and rises run between plateau boundaries.

QualASSESS uses heuristics aimed at robustness rather than precision - this is essential because even a few outliers could seriously distort statistics.

Operations on the intensity contour begin with an averaging procedure which smooths out fluctuations within the duration of a typical syllable (about 150ms). Secondary smoothing removes 'zig-zags' which break a trend that is present on both sides of the 'zig-zag'. Inflections and plateaux are then found straightforwardly. A preliminary identification of pauses is also made, using the principle that levels associated with pauses are likely to be much more common than the levels just above them.

Fricative bursts are considered next. Slices are tentatively classed as fricative if the ratio of energy above 2.5kHz to that below 625Hz exceeds an empirically set threshold. Then a second pass checks for slices which are more likely to be pauses, using tests based on their duration, spectrum, and energy level. Evidence from all three is summed to decide whether slices are reassigned as pauses.

Pause finding continues by finding the average spectrum for slices which have been classified as pauses to date. All slices are compared to the resulting pause spectrum, and those with similar profiles are reclassified as pauses. So are short periods flanked by pauses on both sides.

The next task is to obtain robust estimates of voice pitch. To limit the impact of dubious pitch estimates, data

from QuantASSESS are completely ignored if they fall within a pause or a fricative burst, or exceed reasonable pitch limits. The variance of the pitch estimate is also calculated, and low confidence is assigned to data in periods with high variance. A 'snake' is then fitted to the pitch estimates. This is modelled on an elastic rope, stretched across the sample and pulled towards each data point by a spring whose strength reflects confidence in the estimate. The snake settles to a stable shape in an iterative process driven by the springs and the elasticity of the rope. This provides a robust estimate of the pitch contour. Peaks, troughs, and plateaux are then identified.

Transitions are now used to partition spectral information.

Three main average spectra are created - for fricatives, for pauses, and the 'main spectrum' for all other periods. Sub-spectra are also constructed for two significant types of episode, fricative bursts and peaks in the intensity contour. Peaks are considered as the best available approximation to vowel centres.

Two summaries are produced for each type. One sums the energy in each filter position: the other sums energy squared. Sums of squares are used later to calculate the variance of energy at each position. Describing variance for each filter gives a spectrum-like result which shows the variability of energy at each filter rather than its intensity. These variance spectra are designed to indicate whether episodes such as fricative bursts or vowels show normal or reduced variation from moment to moment.

The final part of QualASSESS deals with pitch again. QualASSESS uses a crude, but serviceable definition of tune boundaries: they are defined by pauses lasting longer than 150ms. It summarises the shape of each tune by fitting a curve with two components, a straight line and a quadratic (U-shaped) curve, with its centre on the midpoint of the tune.

STATISTICAL SUMMARY

SummASSESS links QualASSESS files derived from a single passage. It also has a normalising function. For each passage it reads a 'calibration' file to set a scale for intensity measurements. If the calibration file contains a tone of known intensity, this can be used to give true dB values. If not, median intensity can be set to a default value (usually 60dB).

SumASSESS generates statistical descriptions in two forms: a graphic summary showing the pitch and intensity contours, fricative bursts, and subspectra; and a statistical table consisting of three main blocks, dealing with properties of relatively high-order units; properties of spectra; and properties of the main contours, intensity and pitch.

Three sub-blocks describe high order units.

(1) Tunes. Tables give the number of tunes in a passage, and the average and standard deviation of several properties - duration, parameters of the fitted curves, number of inflections per tune, maximum and minimum pitch values within a tune, and the slope and the duration of opening and closing segments.

(2) 'Sound blocks'. These are stretches of intensity contour between two pauses. Mean and standard deviation are given for number of peaks per block, maximum height, and duration: and also for the duration, height, and slope of segments which open and close sound blocks.

(3) Frication. This uses measures based directly on the fricative spectrum and descriptors which deal with fricative bursts, including number of bursts and mean and standard deviation of burst properties such as duration, amount of energy, midpoint of energy, and spread of energy across the fricative region.

The spectrum section deals with five spectra, namely main; fricative burst average and variance; and intensity peak average and variance. For each, a matrix of measures describes four broad bands, covering the regions which tend to contain F0, F1, F2 and frication. For each band, SummASSESS specifies total activity, average activity per filter channel, variation between filter channels within the region, and centralisation or spread of activity across the region. The shape of each spectrum is summarised by the slope of a fitted line, and measures of its midpoint. Together these capture most properties of the spectrum that previous work suggests may be relevant [3], [11].

Contour measures form two parallel matrices for intensity and pitch. The basic unit is a row which specifies a feature's frequency, the mean and standard deviation of its value, and non-parametric measures - median, 10% point, 90% point, and inter-quartile range. Each matrix covers the following features: magnitude (dB for intensity, Hz for F0) for all points, peaks, troughs, rises, and falls: and duration measures for pauses, rises, falls, and plateaux.

PERFORMANCE

Results have been obtained from ASSESS in two major projects and a subsidiary one. They confirm the potential of the approach. The first major project [14], studied speech in 75 deafened speakers before and after cochlear implantation, plus 51 controls. It confirms and extends earlier observations on pre-implantation speech, and provides objective evidence that implantation produces changes - not all in the right direction. The second major project deals with the speech of 72 schizophrenic patients. Preliminary results show differences between them and controls, in a wider range of attributes than previous studies have considered. The subsidiary project attached to that studies vocal expression of emotion. Results show distinctions between passages expressing different emotions in spectral balance, range of pitch movement, timing of pitch movement, timing of intensity changes, and intensity distribution [15].

In none of these areas does ASSESS offer a complete analysis. Rather it is a broad brush tool, able to examine substantial bodies of speech and to establish that effects of certain general types exist. That seems a useful addition to the repertoire of phonetic science.

ACKNOWLEDGEMENT is due to Dr. D. Howard for pitch extraction algorithms.

REFERENCES

[1] Cowie, R. & Douglas-Cowie, E. (1983), "Speech in postlingual deafness", in M. Lutman and M. Haggard (eds.), *Hearing Science and Hearing Disorders*, London: Academic Press, pp. 183-230.

[2] Goehl, H. & Kaufman, D. (1986), "The real thing: a reply to Cowie, Douglas-Cowie & Stewart", *J. of Speech & Hearing Disorders* vol 51, pp.185-187

[3] Cowie, R. & Douglas-Cowie, E. (1992), *Postlingually acquired deafness: speech deterioration and the wider consequences*, Berlin: Mouton de Gruyter.

[4] Cowie, R., Douglas-Cowie, E. & Rahilly, J. (1991), "Instrumental measures of abnormalities in deafened speech", *Proc 12th ICPhS*, Aix-en-Provence, pp. 350-353.

[5] Rahilly, J. (1991), "Intonation patterns in postlingually deafened and normal hearing people in Belfast", Ph.D dissertation, Queen's University Belfast.

[6] Douglas-Cowie, E., Cowie, R. & Rahilly, J. (1994), "The social distribution of intonation patterns in Belfast", in J. Windsor-Lewis (ed.), *Studies in General and English Phonetics, In Honour of J.D. O'Connor*, London: Routledge, pp. 180-186.

[7] Leff, J. & Abberton, E. (1981), "Voice pitch measurements in schizophrenia and depression", *Psychol. Medicine*, vol. 11, pp. 849-852.

[8] Frick, R. (1985), "Communicating emotion: the role of prosodic features", *Psych. Bulletin*, vol. 97, pp. 412-419.

[9] Scherer, K. (1986), "Vocal affect expression: a review and a model for future research", *Psych. Bulletin*, vol. 99, pp. 143-165.

[10] Murray, I. & Arnott, J. (1993), "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *J. Acoust. Soc. Am.*, 93 (2), pp. 1097-1108.

[11] Knapp, M. (1972), *Nonverbal communication in human interaction*, New York: Holt, Rinehart and Winston.

[12] Hammarberg, B., Fritzell, B., Gauffin, J. et al. (1980), "Perceptual and acoustic correlates of abnormal voice qualities",*Acta Otolaryng* 90 pp.441-451

[13] Andreasen, N., Alphert, M. & Merrill, J. (1981), "Acoustic analysis: an objective measure of affective flattening", *Arch. Gen. Psychiatry*, vol 38, 281-285.

[14] Cowie, R., Douglas-Cowie, E., Sawey, M. et al. (1995), "The effects of cochlear implants on speech production in postlingually acquired deafness", *Proc.13th ICPhS*, Stockholm.

[15] McGilloway, S., Cowie, R. & Douglas-Cowie, E. (1995), "Prosodic signs of emotion in speech: preliminary results from a new technique for automatic statistical analysis", *Proc.13th ICPhS*, Stockholm.