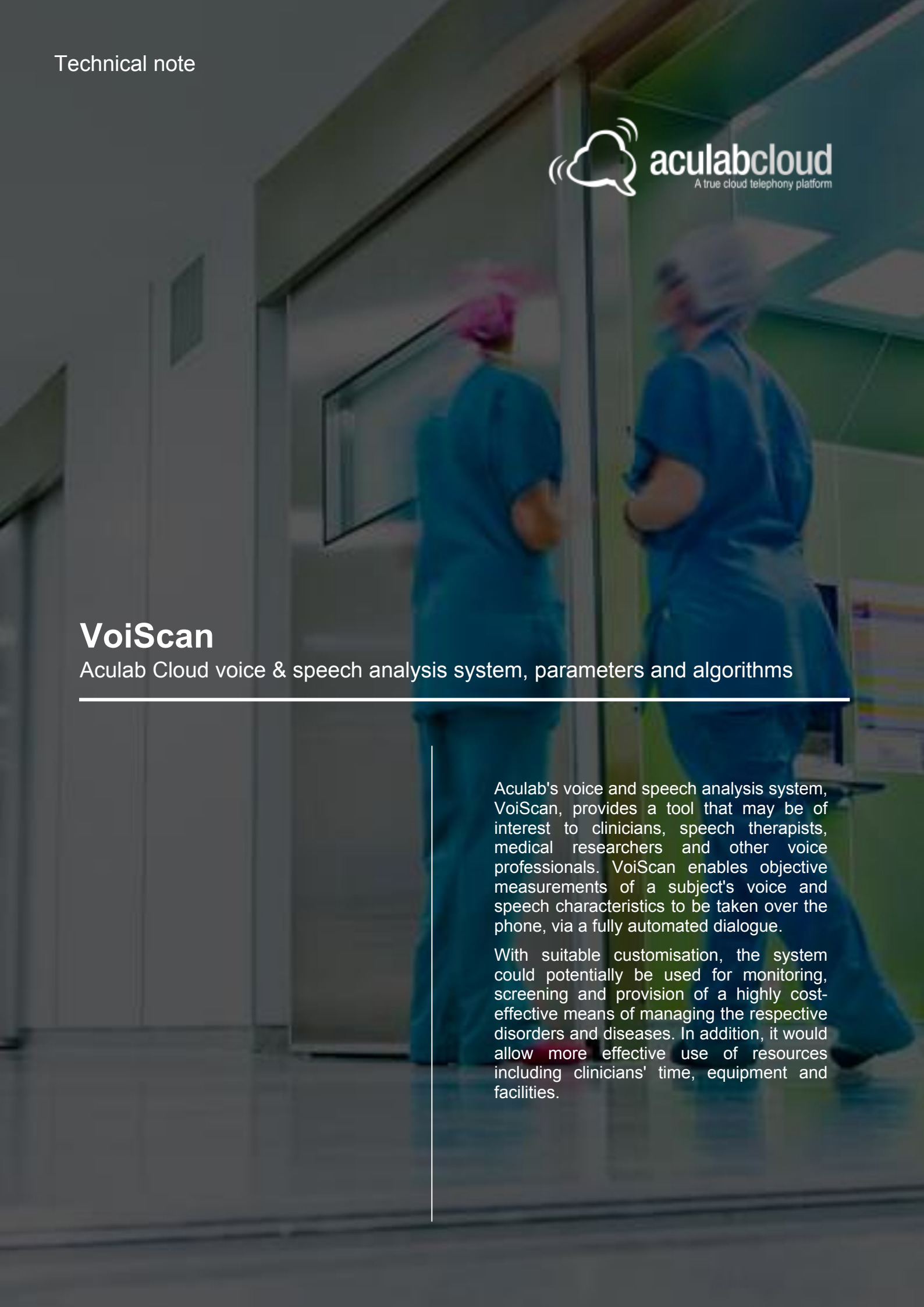
Technical note



VoiScan

Aculab Cloud voice & speech analysis system, parameters and algorithms

Aculab's voice and speech analysis system, VoiScan, provides a tool that may be of interest to clinicians, speech therapists, medical researchers and other voice professionals. VoiScan enables objective measurements of a subject's voice and speech characteristics to be taken over the phone, via a fully automated dialogue.

With suitable customisation, the system could potentially be used for monitoring, screening and provision of a highly cost-effective means of managing the respective disorders and diseases. In addition, it would allow more effective use of resources including clinicians' time, equipment and facilities.

Introduction

Many diseases and medical conditions affect a subject's voice and the patterns of their speech. The assessment and diagnosis of these conditions generally involves attendance at specialist clinics where speech and language therapists, or other voice specialists, analyse a number of characteristics of the speech. Their analysis is largely subjective and requires significant levels of training and expertise on the part of the clinician. A detailed history of objective measurements would allow clinicians to make a more informed and accurate assessment, but this can be time-consuming and expensive since regular clinical visits would be required.

Making audio recordings over the telephone can provide a large cost saving, and simultaneously minimise the disruption to both the subject and the clinician. Telephone-quality speech contains a wealth of information, not only allowing for a caller's speech to be understood, but also for individual speaker characteristics to be identified, and for many abnormalities to be detected and quantified.

Calls can be scheduled as frequently as necessary, and the system can include additional communication between clinician and patient, such as confirmation of the patient's availability for clinical appointments or prompting the patient to perform any self-medication or other actions that may be required.

Demo

A simple demonstration of VoiScan has been set up. In this demonstration, a subject's responses to a short series of spoken prompts is recorded, analysed, and a report returned to them via email. In a practical system, the email would be sent to a clinician, therapist or other speech or voice specialist, who would then interpret the results and compare them with the subject's history, before deciding whether a clinical visit was warranted.

Although not a part of this demo, it is possible for the Aculab system to keep a record of previous analyses and present the whole history to the clinician once sufficient data has been accumulated.

The demo procedure is summarised as follows:

1. Send an SMS text message consisting of your email address to:
   +44 7441 909559 (UK) or
   +1 321 300 2424 (USA).
2. Wait for an SMS reply which will confirm your registration and provide instructions for the next step.
3. Call the number specified in the SMS instructions.
4. Follow the instructions you hear. You will be asked to memorise a number, say "aah" for 5 to 10 seconds, and then say the memorised number. Finally you will be asked to repeat two sentences from a standard diagnostic text (the Grandfather passage), and the call will end.
5. Aculab's automatic voice and speech analysis system analyses your voice, compares it with similar recordings from thousands of healthy adult speakers, and emails you the results.

The email will include the analysis results as a large number of speech and voice parameters, which quantify many of the characteristics typically used by speech therapists to assist them with clinical diagnoses. These parameters are described in detail in the next section. The email also includes graphical summaries of some of the more informative results as well as a waveform plot, spectrograms, and the audio recordings used for the analysis.

The parameters are also tabulated as numerical quantities, together with the corresponding percentiles of healthy adult speakers in the population. Any parameters below the 5th percentile, or above the 95th, are flagged as being “unusual”, but it is very common for an individual set of parameters to yield as few as 80% of parameters within the “normal” range.

***Note that this system is for demonstration purposes only. The results should not be taken as an indication of any health issue. In addition, the demo only presents a subset of Aculab’s voice and speech analysis system’s capabilities, as described in the next section. By using this demo you agree that Aculab may record and store your voice and use it to improve it’s services, such as this demo. To protect your privacy, we take steps to de-identify your data and keep it secure. We won’t publish your data or let other people use it.**

Demo parameters and algorithms

The fundamental parameters calculated by the VoiScan system have been designed to be extracted from unconstrained telephone recordings of subjects’ voices.

Aculab has drawn on its extensive experience in real-world telephone systems to ensure that the specifications of the parameters, and the algorithms used to calculate them, provide the fullest and most accurate information regarding each subject’s voice and speech.

The individual parameters measured in this demo are defined below:

Cognition

The demo performs a simple memory test using Aculab’s VoiScan system.

The subject is asked to memorise a three digit number and repeat it after performing another unrelated task. If required, much more thorough tests can be conducted, using speech or DTMF key presses to answer questions and assess response times, characterising some aspects of cognitive function.

Voice source parameters

The parameters listed in this section can be calculated from both sustained phonations (steady vowel sounds) and the voiced (vowel-like) parts of natural speech. In the former case, the values are estimated from the longest continuous period of voicing identified in the audio recording. Otherwise they reflect an average over all the voiced phonemes in the whole of the audio recording.

Most of these parameters characterise the voice source (the vocal folds and the other structures within the larynx), but they can be indicative of both physiological and neurological disorders.

Response time (s)

The elapsed time between the end of the spoken prompt and the start of detected speech.

If this is unusually long, then the speaker may have been subject to distraction, may not have heard the instructions for some reason, or have other cognitive issues.

Active duration (s)

The total duration from the start to the end of detected speech.

If this is very short or very long, the speech may have been mis-detected or the speaker may not have complied with the instructions. The audio recordings may need to be examined to identify the reason.

Total breaks

The number of pitch discontinuities in the voicing of the detected speech.

A large number of discontinuities during sustained phonation may simply indicate a poor quality mobile phone line, but can also indicate medical issues such as dysphonia.

Analysed interval (s)

The duration of the segment of the signal subjected to detailed analysis.

For sustained phonation this represents the longest continuously-voiced interval. For other analyses, it is the total duration of all the voiced segments. If this figure is too small, the remaining parameters may be unreliable and the test should be repeated.

Pitch (Hz)

The mean fundamental frequency of the voiced speech.

This is typically in the range 80 to 200 Hz for adult males, and 150 to 340 Hz for adult females. An unusually low pitch is often associated with increased levels of Creakiness (below), even in healthy speakers, while an unusually high pitch can adversely affect the resolution of formant analysis. Thus it is important to check the Pitch parameter before drawing any conclusions regarding Creakiness or the Vocal Tract Parameters.

Pitch (standard deviation, semitones)

A measure of the range of pitch frequencies observed during the analysed segment.

A low value during sustained phonation indicates that the speaker has good control over their vocal folds, whereas a low value during natural speech indicates that the speech is literally monotonous, and the speaker may have been unable or unmotivated to produce a natural inflection for some reason (possibly a neurological disorder or depression).

Jitter (percent)

The short-term variation in the timing of the vocal folds' vibrations.

High values indicate irregularity in the vocal folds' movement which can be due to many factors, including both physical and neurological conditions.

Shimmer (dB)

The short-term variation in the amplitude of the vocal folds' vibrations.

High values indicate irregularity in the air flow between the vocal folds, which can be due to a range of conditions affecting the control of the muscles within the larynx, amongst other things.

Noise-to-harmonics ratio (dB)

The energy in any noise-like components (turbulence) relative to the periodic components (harmonics) during voiced speech.

A clear singing voice should exhibit a low Noise-to-Harmonics Ratio (NHR), i.e. a very negative value when expressed in dB. Conversely, a less negative value indicates a roughness to the voice, such as that caused by laryngitis or other diseases of the larynx.

Breathiness

An estimate of the perceived breathiness of the voice, calculated using a telephone-optimised version of the Fukazawa Breathiness Index.

This is also an indicator of laryngeal pathologies, but designed to reflect similar aspects of the speech to those identified perceptually by clinicians. The original method has been enhanced using modern techniques for high-resolution time-frequency analysis.

Creakiness

A numerical parameter which correlates well with perceived creakiness of the voice, but which has been optimised for robust analysis of telephone speech.

A high value indicates that there is a periodic irregularity in the voice waveform, such as diplophonia (which can be caused by unilateral vocal fold paralysis or by cysts on the vocal folds).

Vocal tract parameters

These parameters can be calculated from both sustained phonations (steady vowel sounds) and natural speech. In the latter case, the values returned reflect an average over all the phonemes in the audio recording. They are based on estimates of acoustical characteristics of the vocal tract, specifically the resonant frequencies (the “formants”) and the trough between the first two such formants.

F1 (Hz)

The average frequency of the first vocal tract resonance (formant).

This allows detection of abnormalities in the articulation of steady vowels (typically “aa”, “ee” or “oo”). An unusually high or low value can simply indicate an unusual accent, or possibly a medical condition such as Apraxia of Speech (AoS).

A1 (relative, dB)

The energy of the first formant, expressed as a proportion of the first three formants.

An unusual value may indicate a medical condition normally associated with Breathiness, but it can also be caused by a poorly-located microphone or other electro-acoustic problem with the telephone.

F1 sharpness (percent)

The sharpness of the first formant resonance.

This parameter should be highest for trained classical singers who are practiced in matching the pitch of their voice to the resonant frequencies of the vocal tract. It gives an indication of the perceived clarity of the vowel being analysed.

F1 variation (semitones)

The amount of short-term variation in the first formant frequency.
During sustained phonation, this value should be small, indicating a stable vowel sound, but during fluent speech, significant variation is normal and unusually small variation can be symptomatic of poor articulation.

A1 variation (dB)

The amount of variation in the amplitude of the first formant.

The same comments as for F1 Variation, above, apply here too.

F2 (Hz)

The average frequency of the second formant.

This allows detection of abnormalities in the articulation of steady vowels (typically “aa”, “ee” or “oo”). An unusually high or low value can simply indicate an unusual accent, or possibly a medical condition such as Apraxia of Speech (AoS).

A2 (relative, dB)

The energy of the second formant, expressed as a proportion of the first three formants. The same comments as for A1, above, apply here too.

F2 sharpness (percent)

The sharpness of the second formant resonance.

The same comments as for F1 Sharpness, above, apply here too.

F2 variation (semitones)

The amount of short-term variation in the second formant frequency.
As for F1 Variation, this value should be small during sustained phonation, indicating an unchanging vowel sound. However during normal fluent speech, the second formant normally covers a very wide range and a small F2 Variation suggests under-articulation of the vowel sounds. Thus it can be an indication of significant speech impairment.

A2 variation (dB)

The amount of variation in the amplitude of the second formant.

The same comments as for A1 Variation, above, apply here too.

F1,F2 trough frequency (proximity to F2, percent)

The relative frequency of the trough between the first two formants.
A value of zero corresponds to the trough being adjacent to F1, while a value of 100% corresponds to it lying next to F2. Most non-nasalised sounds produce a value in the region of 50%, but a value at either end of the range can indicate the presence of unusual nasalisation.

F1,F2 prominence (dB)

The normalised energy of the first two formants relative to that of the trough between them.

This is indirectly related to the F1 and F2 Sharpness measures.

F1,F2 trough sharpness (percent)

The sharpness of the trough between the first two formants.

A high value can indicate unusual levels of nasalisation.

F1,F2 trough frequency variation (semitones)

The amount of short-term variation in the F1,F2 Trough Frequency.

If this is large then the estimate of F1,F2 Trough Frequency may have been affected by background noise or other factors.

F1,F2 prominence variation (dB)

The short-term variation in the F1,F2 Prominence.

Indicates how smoothly the voiced sounds are articulated, and would be expected to be lower in trained classical singers.

F3 (Hz)

The average frequency of the third formant.

This is normally a fixed value for any given speaker, and can be used to aid interpretation of the first two formant frequencies, F1 and F2. If F3 is very different from previous values observed for the same speaker, then the recordings need to be examined for anomalies.

A3 (relative, dB)

The energy of the third formant, expressed as a proportion of the first three formants.

This is normally lower than A1 and A2, but if it is too low, the estimated F3 may be inaccurate.

Articulatory dynamics parameters

These parameters are aimed at the analysis of natural speech, and characterise the articulation and control of the upper vocal tract. However, they can also be estimated from sustained phonations in order to identify non-compliant subjects, or technical problems with the recordings.

Each of these parameters is calculated without explicit identification of formants, using telephone-robust algorithms which operate on the broad time-frequency structure of the speech signal.

Preliminary experiments have shown that these parameters can contribute to the automated analysis of disordered speech, but they should be considered experimental at this time, since they are not clinically proven.

Frequency-time uncertainty

A telephone-robust version of the Hirschman Uncertainty.

This provides a measure of how diffuse the energy in a spectrogram is. A large value indicates that the energy is smoothly spread across large regions of the spectrogram and the recorded speech is poorly articulated, either because of damped vocal tract resonance or imprecise voice source transitions.

Frequency-time orientation

The frequency-domain entropy of the spectrogram relative to the time-domain entropy.

If the energy in the spectrogram is spread across a wide range of frequencies, but concentrated in short time intervals (e.g. stutter-like sounds: “p-p-p-...”, “t-t-t-...”, or “k-k-k...”) this value will be clearly positive. Conversely, if the energy is spread over a long period of time, but concentrated in a small range of frequencies (e.g. when whistling or pressing the digits of a DTMF telephone keypad), this value will be clearly negative.

Dynamics

The average short-term direction-change of the frequency components in the speech.

This value is close to zero during sustained phonation, but generally negative during natural speech. In that case, a low (i.e. more negative) value indicates that the formants change direction less often, which can be associated with a slow rate of speaking. Again, during natural speech, a high value indicates frequent changes in direction and a fast rate of speaking.

Dynamics (standard deviation)

The standard deviation of the Dynamics parameter over the duration of the speech.

Since the Dynamics parameter gives an indication of speaking rate, this value reflects the variation in speaking rate during an utterance. Thus it yields information regarding prosody and the functioning of the speaker’s linguistic processes.

Dynamics (skewness)

The sample skewness of the Dynamics parameter over the duration of the speech.

This is derived from the Dynamics parameter, and has been found to be correlated with the presence of some speech and voice disorders. A low (negative) value indicates that the Dynamics parameter is generally low with brief periods of increased activity, i.e. a slow speaking rate with occasional rapid articulation. This may indicate, for example, a speaker with good motor control and unusually precise speech.

Dynamics (kurtosis)

The sample excess kurtosis of the Dynamics parameter over the duration of the speech.

This is derived from the Dynamics parameter, and has been found to be correlated with the presence of some speech and voice disorders. It indicates the presence of outliers (unexpectedly extreme values) in the Dynamics parameter.