Chapter 2
Speech Production and Perception

Abstract Certain specific characteristics of speech are known to be particularly useful in diagnosing speech disorders by acoustic (perceptual) and instrumental methods. The most widely cited of these are described in this chapter, along with some comments as to their suitability for use in automated systems. Some of these features can be characterised by relatively simple signal processing operations, while others would ideally require a realistic model of the higher levels of neurological processing, including cognition. It is observed that even experts who come to the same ultimate decision regarding diagnosis, often differ in their assessment of individual speech characteristics. The difficulties of quantifying prosody and accurately identifying pitch epochs are highlighted, because of their importance in human perception of speech disorders.

Keywords Characteristics of disordered speech · Objective assessment of vocal characteristics · Precision of articulation · Subjectivity of perception

The higher levels of neural processing involved in both speech production and speech perception are still only partially understood. There is even some degree of controversy regarding the detailed mechanical functioning of the peripheral auditory system (Shera and Olson 2011), despite having been the subject of research for many years.

According to Duffy (2000), "speech is the most complex of innately acquired human motor skills, an activity characterised in normal adults by the production of about 14 distinguishable sounds per second through the coordinated actions of about 100 muscles innervated by multiple cranial and spinal nerves". Furthermore, speech production involves great temporal precision (in the region of 10 ms), while some parts of the peripheral auditory system have been found to exhibit an even greater resolution—in the region of 10 μs (Fuchs 2005).

The complexity of these processes, and the fine temporal resolution they utilise makes precise diagnosis of the location of neural damage extremely difficult, and

L. Baghai-Ravary and S. W. Beet, *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*, SpringerBriefs in Speech Technology,
DOI: 10.1007/978-1-4614-4574-6_2,

only some of the features proposed for use in diagnostic aids should be expected to yield useful and reliable information.

Masaki (2010) provides a thorough and well thought-out description of the field and discusses development of the voice production structures. She mentions that some are not fully developed, even until the age of 25, so it should be borne in mind that voice production is not a stable and fixed process, even for one individual, and the effect of some pathologies on the voice will be more or less profound at different stages of vocal development.

2.1 Articulation

At the core of the definition of the term "dysarthria", is mis-articulation of the sounds in spoken words. It can manifest itself in all aspects of speech production: phonation, resonance, articulation and respiration. As such, analysis of the accuracy of articulation is clearly important to diagnosis and telemonitoring.

The accuracy of articulation can be affected by many factors—both neurological (as in cerebral palsy, for example, where the precision, speed and variety of the sounds is often restricted) and physical (due to factors such as the health of the muscles used to produce the sounds). The situation is somewhat complicated since long-term neurological conditions can adversely affect physical attributes such as muscle tonicity. In fluent speech, articulation can be abnormally restricted (if the muscles are hypertonic, as in spastic cerebral palsy) or abnormally large (if they are hypotonic). The various articulators can be affected to different degrees depending on the specific condition. Articulation can also be affected by physical or neurological damage in the mechanisms associated with speech production. Even disturbances affecting the subject's perception of their own speech can cause abnormal articulation.

Whenever articulation is disturbed, it generally becomes more difficult to understand the speech, so a speech and language therapist will generally grade the intelligibility (see below) as part of the assessment. A more detailed analysis of the exact form of the disturbance can be very time-consuming.

2.2 Phonation

The term "phonation" refers to periodic modulations of the air pressure from the lungs by the opening and closing of the vocal folds. There are a number of causes of atypical phonation, both physical and neurological in origin.

Phonation can become irregular in terms of pitch ("jitter"), or in terms of amplitude ("shimmer"), or it can be polluted with aperiodic noise, caused by turbulence around a constriction in the airway, for example. The presence of aperiodic noise is conventionally quantified via the harmonic-to-noise ratio, HNR. It is also very common for the maximum duration of sustained phonation to be restricted in pathological speech. All of these factors are routinely assessed using

computerised analysis of the speech waveform, although they can only be evaluated accurately and repeatably from sustained vowel phonations.

For the purposes of assessing phonation, it is common to analyse sustained production of the /a/ vowel (/a/ as in "father"), and indeed for automatic systems, it has been shown that this specific vowel offers noticeably better accuracy than others (Henríquez et al. 2009).

2.3 Voicing

A number of structures can be utilised for the production of the acoustic pressure wave that excites the resonances of the vocal tract. This excitation signal can be one or more of periodic, transient, or stochastic (fricative) in nature. Any abnormality can be heard in the resulting speech, and the type of voicing can usually be identified quite reliably by automatic methods.

A number of speech disorders can manifest themselves as this type of abnormality. For example, phonation may be present during what should be fricative sounds, or there may be errors which effectively realise a voiced stop as unvoiced, or vice versa.

2.4 Resonance

When voiced speech is produced, energy may be coupled into the nasal cavity, introducing an additional resonance and anti-resonance into the overall vocal tract response. Perceptual assessments of speech quality attempt to quantify the effectiveness of the nasal coupling by gauging the degree of "nasal resonance".

Some tests additionally assess the "oral resonance" of the speech, which is a measure of how far "back in the throat" the voice is perceived as being.

Both oral and nasal resonance is usually assessed from read passages containing both nasal and non-nasal sounds. The presence or absence of nasal resonance can be determined quite easily, especially in context (the onset of nasal sounds is clearly visible in a spectrogram, for example) but oral resonance is difficult to quantify precisely, and thus difficult to analyse automatically.

2.5 Prosody

The intonation, stress, and rhythm of speech, i.e. the prosody, contains a disproportionate amount of high-level non-verbal information, much of which is used socially, and as such, is adversely affected in a number of neurological conditions, especially those which affect social interaction. Prosody is also readily disrupted by problems of respiration. Both the pitch and the intensity of the speech can become more difficult to control if the supply of air to the vocal folds is too low or too variable.

In terms of the analysis of prosody, van Santen et al. (2009) suggests that the data produced by commercial software such as the KayPENTAX® Multi-Dimensional Voice Program, MDVP™, is only marginally relevant to prosody, even though prosody is generally believed to be key in perceptual/acoustic assessment of disorders. This is because the phonetician's concept of prosody is defined at a more abstract level, concerned with concepts such as "stress", "intonation" and "rhythm", which can be assessed in terms of "naturalness" and other perceived characteristics, but which cannot be defined precisely via mathematical equations.

The data provided by automatic systems can only represent statistics of the signal, either integrated over a varied example of speech (e.g. the range and stability of pitch during a sentence), or evaluated independently over individual speech sounds (jitter or shimmer measured during sustained phonation, voice-onset-time (VOT) for stop consonants, etc.).

van Santen et al. (2009) goes on to quote others saying that dialect-differences can be as large as language-differences in prosody, and accents can cause problems for both human and automatic assessment. They also discuss the advantages human assessment can offer because of the ability to adapt to accent, voice quality, speaking style, etc.

It can be inferred from Kent (1996) that human assessment cannot be relied on to provide a benchmark for the development of automatic prosodic analysis: "the acoustic correlates of prosody are perceptually much too complex to be fully categorized into items by humans—they cannot be reliably judged by humans who have subjective opinions". That is to say that there is currently no way to characterise the subtleties of prosody in a tractable mathematical form.

Nonetheless, a number of useful mathematical features related to prosody have been defined, (e.g. Pentland 2007; Schuller et al. 2007), and they can identify at least some gross abnormalities in prosody. Despite such progress in defining a wide set of prosodic features, there is still no clear consensus as to the most effective or efficient features (Ringeval et al. 2010).

2.6 Intelligibility

For many conditions, where a simple measure of its severity is required to monitor response to treatment, for example, a perceptual assessment of intelligibility can be used.

As long as the sounds being produced can be controlled as part of the assessment protocol, intelligibility can be assessed acoustically, although there is evidence that different listeners may produce noticeably different evaluations (McHenry 2011). Indeed in (Ziegler and Zierdt 2008) it was found to be necessary to average the results of 2 or 3 different listeners in order to produce a reliable correlation between manual and automatic measures, recorded on a simple percentage scale. When assessment is to be performed by a single listener, a graded assessment is more appropriate (on a 4 or 5 point scale, for example).

References

Duffy JR (2000) Motor speech disorders: clues to neurologic diagnosis. In: Adler CH, Ahlskog JE (eds) Parkinson's disease and movement disorders: diagnosis and treatment guidelines for the practicing physician. Humana Press, Totowa, pp 35–53

Fuchs PA (2005) Time and intensity coding at the hair cell's ribbon synapse. J Physiol 566(1):7–12

Henríquez P, Alonso JB, Ferrer MA, Travieso CM, Godino-Llorente JI, Díaz-de-María F (2009) Characterization of healthy and pathological voice through measures based on nonlinear dynamics. IEEE Trans Audio Speech Lang Process 17(6):1186–1195

Kent RD (1996) Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. Am J Speech-Lang Pathol 5(3):7–23

Masaki A (2010) Optimizing acoustic and perceptual assessment of voice quality in children with vocal nodules. PhD thesis, Harvard-MIT Health Sciences and Technology

McHenry M (2011) An exploration of listener variability in intelligibility judgments. Am J Speech-Lang Pathol 20:119–123. doi:10.1044/1058-0360(2010/10-0059

Pentland A (2007) Social signal processing. IEEE Signal Process Mag 24(4):108–111

Ringeval F, Demouy J, Szaszák G, Chetouani M, Robel L, Xavier J, Cohen D, Plaza M (2010) Automatic intonation recognition for the prosodic assessment of language-impaired children. IEEE Trans Audio Speech Lang Process 19(5):1328–1342. doi:10.1109/TASL.2010.2090147

Schuller B, Batliner A, Seppi D, Steidl S, Vogt T, Wagner J, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V (2007) The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proceedings of INTERSPEECH-2007, pp 2253–2256

Shera CA, Olson ES (eds) (2011) What fire is in mine ears: progress in auditory biomechanics. In: Proceedings of 11th international mechanics of hearing workshop. American Institute of Physics, Melville

van Santen JPH, Prud'hommeaux ET, Black LM (2009) Automated assessment of prosody production. Speech Commun 51(11):1082–1097. doi:10.1016/j.specom.2009.04.007

Ziegler W, Zierdt A (2008) Telediagnostic assessment of intelligibility in dysarthria: a pilot investigation of MVP-online. J Commun Disord 41(6):553–577

http://www.springer.com/978-1-4614-4573-9

Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders

Baghai-Ravary, L.; Beet, S.W.

2013, VIII, 70 p. 9 illus., Softcover
ISBN: 978-1-4614-4573-9