DRAFT V01
2020-10-31

Phonetics for Fieldwork

Dafydd Gibbon

Contents

1 Phonetic methods

This contribution is based on the assumption that many fieldworkers, not only in linguistics but also in neighbouring disciplines, do not necessarily have detailed phonetic training, and yet an ability to distinguish and often produce speech sounds which differ from those in their native language is required: interpreting meanings requires understanding sounds. The opposite is not necessarily true, interestingly: one experiences words as sounds without meaning (except for associations implied by the context) when one comes across previously unknown words. A trained phonetician will find some of the explanations rather basic. On the other hand, an ethnologically oriented researcher who is mainly interested in narrative data will find some of the information to be overkill. Nevertheless, it is customary to provide access to recorded data to the originating speech community as ‘payback’ or simply common courtesy, and for the same reason the quality of the recording should be as high as is reasonably possible. The information in this contribution is designed not only for phonetic and linguistic research but also with this aim in mind. Clearly, a contribution of this kind cannot replace a tutorial. Rather, the aim is to provide systematic pointers to further information, and to add practical fieldwork oriented information which is typically not found in the introductory literature.

As in any other science, the phonetician starts with an intuitive understanding of the domain of speech events and refines this understanding with a broad palette of scientific methods:

1. auditory, visual and tactile skills for identifying and classifying speech events;
2. systematic planning, recording and storage of data;
3. phonological analysis with contrastive segmentation and classification of speech events;
4. structured data acquisition in experimental paradigms;
5. quantitative descriptive statistical analysis of the data;
6. physical measurements of the data with advanced statistical modelling;
7. formal and computational modelling of structural patterns to match the data;
8. evaluation and validation of psycholinguistic and sociolinguistic models with the data;
9. publication of results in theses, books and articles in open access or commercial outlets;
10. materials for community feedback and payback: books, audio, video recordings, software.

All of these methods are shared in various configurations with other sciences, and trans-disciplinary cooperations are quite common: with linguists of other specialisations; with ethnologists, sociologists and psychologists; with computer scientists, software developers and speech engineers; with medical personnel for speech impairment diagnosis and speech therapy; with educators and language planners; with artists and layouters for payback publication. Fieldworking phoneticians typically select several of these methods and cooperations for their work in the field and in later evaluation of their fieldwork results for publication, for the development of applied phonetic materials and devices, and for feedback and service to the community. This contribution concentrates on the first three methodological areas, and looks selectively at data analysis and computational modelling.

2 Auditory, visual and tactile phonetic skills

2.1 Domains of speech

Phonetics is concerned mainly with the physical domains of speech: the two anatomical and physiological domains of speech production and perception, and with the acoustic domain of speech transmission. The acoustic domain depends very much on the ability to perform measurements with dedicated hardware and software instruments, but the research domains of speech production and perception involve class-taught and self-taught skills. The production and perception domains are epistemologically ambiguous: on the one hand, they are domains for observation in local fieldwork partners, and simultaneously they are skill types on which the phonetician is strongly dependent for their own research. The domains will be outlined primarily with the acquisition of phonetic skills in mind.

The introductory literature on phonetics contains detailed descriptions of each of the domains of speech, including the anatomy of speech production and perception. It is sometimes useful to think more abstractly about these processes, however, with the production of speech as an air-driven machine with an oscillator and noise generators for sound production, and an oral and nasal filter system for modifying the speech sounds (Figure 1). The anatomical model for perception would not be too informative, being inaccessible to observation, and consisting of a microphone membrane, a physical impedance transformer and a physical spectral transformation, with almost all the interesting processing taking place in the brain.

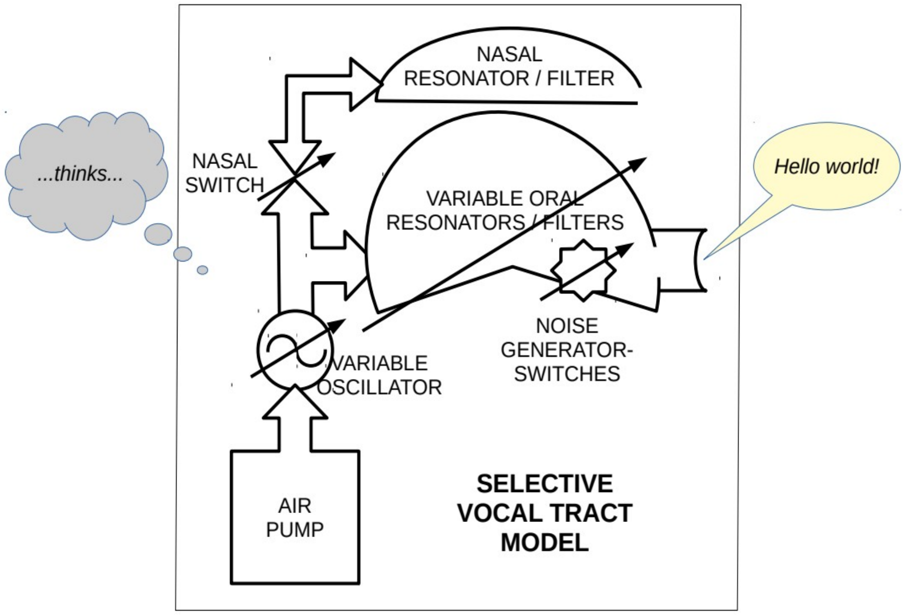


Figure 1: Selective vocal tract model: lungs (air pump), laryngeal source (variable oscillator), nasality (nasal switch, nasal filter-resonator), oral cavity (variable oral filters/resonators), obstruent stops and fricatives (variable noise generator-switches). Variability of parameters is denoted with diagonal arrows.

2.2 Transcription skills

The benchmark reference for speech production and perception skills is the International Phonetic Alphabet (IPA), which has undergone continuous development for one and a half centuries and is still occasionally refined by the International Phonetic Association as details from new languages appear. Use of the internationally recognised IPA in transcription and annotation of the data ensures understanding and reproduceability of the results in future analyses, while informal and orthographic transcription systems have limited scientific value. The organising principle of the IPA is articulatory phonetics: the places and methods of articulation of speech sounds. The IPA is often perceived as a little daunting, at least in comparison with the alphabet of one's native language, but one might point out, after all, that transcription is arguably the only methodology whose fundamental concepts, symbols and their relations can be illustrated on just one page (Figure 10).

Systematic transcriptions of observed data are based on well-defined criteria for segmenting and classifying speech events. Narrow phonetic type of transcription, which is inevitably required in the early stages of fieldwork, represents the full range of sound properties described in the IPA, with consonants and vowels associated with diacritics for the exact articulatory places and manners of production. Broad phonetic transcription … symbols for parallel prosodic speech events such as global pitch patterns associated with the intonation of phrases and sentences, and local pitch accents associated with word and sentence stress (as in English or German), lexical pitch accents (as in Japanese or Swedish), and lexical tones (as in the tone languages of Africa, South-East Asia and the Americas). While local lexically relevant speech events such as consonants, vowels and their properties, and some tones, have standard IPA representations, there exists a relatively wide variety of conventions for representing stress-pitch accents, pitch accents and tones with iconic marks (such as acute, grave, circumflex and caron or háček diacritics), with numbers based on pitch height, or, in phonology, with alphabetic letters and special symbols such as asterisks, hash marks (also known as number or pound signs, similar to the musical sharp sign) and plus or minus signs for locations and boundaries. These prosodic conventions are usually specific to certain phonetic communities in different countries, so no general overview can be given here, and reference to the relevant published literature is called for.

McKinney and McKinney (2017) provide a comprehensive introduction to the acquisition of phonetic skills. **OTHERS**

2.3 Consonants and vowels

There are two main kinds of criterion for defining consonants and vowels and the intermediate classes of liquids and glides or semi-vowels: the classification of sounds in terms of their paradigmatic local physical properties, and the segmentation of speech sounds into syllable, word and phrase patterns as components of a syntagmatic ‘sonority curve’, with consonants at the edges of syllables and vowels at the centre of syllables. Each language has a selection of speech sound types, and there are some preferences for certain speech sounds (such as stop consonants and vowels) over others (such as fricatives or liquids and glides). The main types of speech sound are discussed below; further details may be obtained from the IPA chart.

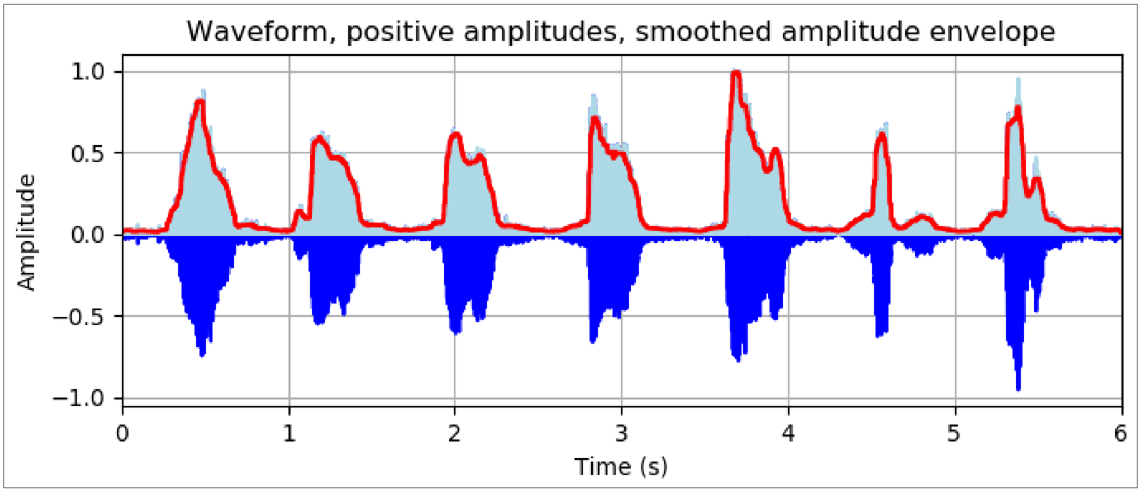


Figure 2: Extraction of the physical correlate of the 'sonority curve': the amplitude modulation envelope (counting from one to seven in English).

The starting point for analysing the speech signal is a representation in the *time domain*, with the *timeline* as independent variable on the *x*-axis and the *amplitude* of the signal waveform on the *y*-axis. Figure 2 shows a time representation of a recording of counting from one to seven in English. The speech signal is 6 s long and the amplitude is sampled at 16 kHz. The resulting 96000 measurements are not individually visible in the compact representation of Figure 2, where the positive amplitude measurements are lighter in colour and the negative amplitude measurements are darker. Linguistically, the important property of this representation is represented by the top line which outlines the positive amplitude variations, and represents the physical counterpart of the 'sonority curve' of alternations between consonants (valleys) and vowels (peaks). The amplitude variation of a speech signal is termed *amplitude modulation*, and the outline shown in Figure 2 above the waveform is the *amplitude modulation envelope*. The amplitude modulation envelope provides an intial set of clues for identifying consonants and vowels. For example, the word *six* [sɩks] is clearly the shortest numeral in the sequence, the result of a short vowel [ɩ] surrounded by voiceless consonants. **The numerous introductions to phonetics on the commercial book market and online provide realistic accounts of the phonetic properties of speech sounds in the articulatory, transmission and perception domains. In the present context, only a very general overview can be given. Above...**

In general, consonants are classified according to manner of articulation, including voicing (voiced and unvoiced, nasal non-nasal) and full or partial constriction of the oral cavity with the paired fixed and movable articulators. Full constriction of the vocal cavity, as in stops (also known as plosives, from the effect produced when opening the vocal cavity) leads briefly to blocking the passage of air, and thus also of sound, through the oral cavity, and is made with the lower lip as movable articulator against the upper lip (as with [p], [b]), the tongue as movable articulator in different positions against the teeth and the roof of the mouth or the throat (as with unvoiced [t], voiced [d], or unvoiced [k], voiced [g]), or with the larynx (as with [ʔ], the glottal stop). With partial constriction, a gap between the articulator pairs remains, so that the passage of air leads to friction and the production of noise, as with the labiodental fricatives [f], [v], with the interdental fricatives [θ], [ð], with the palatal fricatives [ʃ], [ʒ] and with the velar fricatives [x], [ɣ]. The nasal consonants, for example [m], [n], [ŋ] are technically stop consonants with a constriction plus an opening to the nasal cavity, as are the liquids [l] and the rolled, tapped or fricative [r].

The main types of consonant are relatively easy to produce, even if they are not in one's own language, because their articulatory properties are physically defined and can be practiced with clearly defined exercises rather like practising finger movements with musical instruments.

But many consonants which are encountered in fieldwork with local languages in different parts of the world may be rather different. In addition to the main consonant types mentioned above, the other types of consonant can also be constructed with a little thought and effort: clicks (a little like the 'air kiss' for labial clicks, or the 'tut-tut' or 'tst-tst' disapproval sound for alveolar clicks, or some similar sounds used in training pets). The implosive stops are like clicks in that the air moves into the oral cavity through the articulatory movements, not from the lungs, but otherwise place and manner of articulation are like regular stop consonants. The ejective stops are not so easy: one type is rather like gentle spitting, with the lips (labial) or with the tip of the tongue (apical); the k-like ejective is produced analogously. The articulatory mechanism involved is closure and upward movement of the glottis (the gap between the vocal cords); this can be felt by placing a finger on the larynx ('Adam's apple').

The vowel model used by the IPA is the vowel quadrilateral (see Figure 10, "Vowels"), which represents two positional variants of the highest point of the tongue: horizontal (front to back) and vertical (close to open), for example [I] is close and front, [a] is open and front, [u] is open and back and [ɑ] is close and back. The model is a heuristic model rather than a mathematically exact model, in that the vowels are actually determined by the resonance properties of the oral cavity, and these resonance properties can often be reproduced with more than one configuration of the tongue.

A third dimension, lip-rounding, is represented by pairing representations of unrounded vowels such as [i] in a given position with their rounded versions such as [y]. In many languages there is no contrast between rounded and non-rounded vowels, and there is an automatic association of front vowels being unrounded and back vowels being rounded. A fourth dimension of vowel properties is their length as long or short, and as diphthongs (a pair of vowels behaving as a long vowel).

2.4 Acoustic complexity of sounds

The simplest kind of sound is a *sine wave* of a particular frequency, a very 'pure' sound. Natural sounds, in particular speech sounds are much more complex and consist of sounds of many different frequencies added together. In order to inspect these different frequencies a *frequency domain* analysis is used: small segments of the time domain representation are transformed into a different representation in which the *x*-axis represents frequency, and the *y*-axis represents the amplitudes of each frequency in the signal. There are two main kinds of complexity in speech sounds: noise and resonance. Noise consists of a random mix of frequencies in a particular frequency band. For instance, in the recording used here for illustration, the fricative noise of [s] as in *six* is a fairly well defined higher frequency band of random frequencies, while [θ] in *three* has more diffuse noise over a wider frequency band (Figure 3).

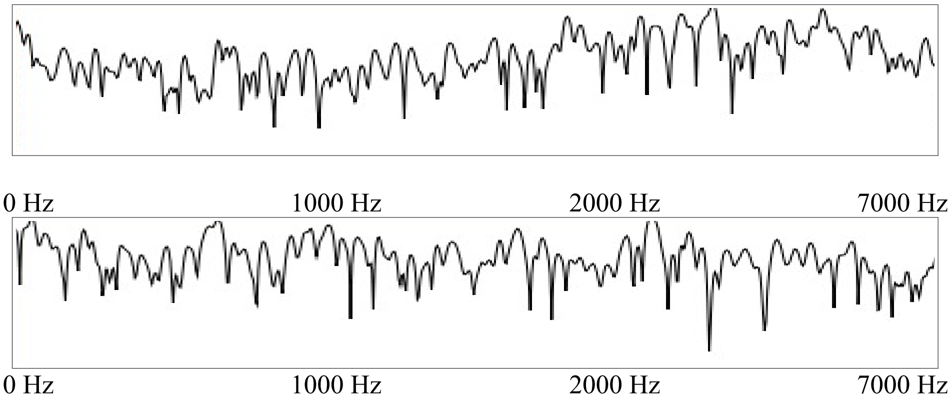


Figure 3: Spectra showing noise patterns, for a male voice (Praat spectral slice).
Top: spectrum of [s] *0...7 kHz, with high frequency noise.*
Lower: spectrum of [θ] *0...7 kHz, with diffuse noise.*

The resonant sounds, including all voiced consonants and vowels, semivowels, nasals, glides,*fundamental frequency* have a different structure. The basic structure represents the complex source oscillation of the vocal cords in the larynx, at the *fundamental frequency*, *F0*, and *harmonics* or *overtones* of *F0*, which are integer multiples of this frequency, at *F0* × *n*. If the fundamental frequency is 100 Hz, the second harmonic is at 200 Hz, followed by 300 Hz, 400 Hz, etc.

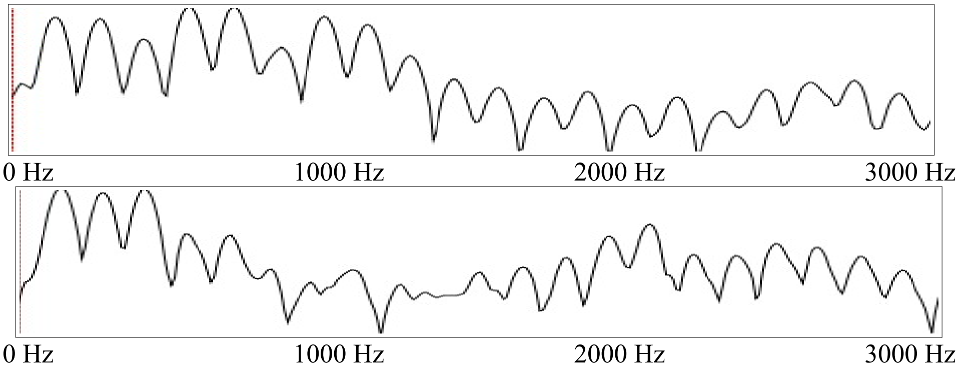


Figure 4: Spectra from [ɑɩ] *in* five*, for a male voice, showing harmonics and formants (Praat spectral slice).*
Upper: [ɑ] *0...3 kHz, formants about 650 Hz and 1.1 kHz..*
Lower: [ɩ] *0...3 kHz, formants 550 Hz and 2 kHz.*

For example, in Figure 4 the fundamental frequency is 145 Hz at the beginning of the diphthong [ɑɩ], on [ɑ] and 139 at the end, on [ɩ]. For [ɑ], 19 harmonics can be counted in the interval

0...3000 Hz, the highest at 145✕19 = 2755 Hz. In the spectrum of [ɩ], 20 harmonics can be counted, in the interval 0...3000 Hz, the highest at 139 ✕ 20 Hz = 2780 Hz. The difference between frequencies of neighbouring harmonics is exactly the same as the fundamental frequency, and this property is used by some algorithms which calculate *F0* in the frequency domain, that is, based on a spectral analysis.

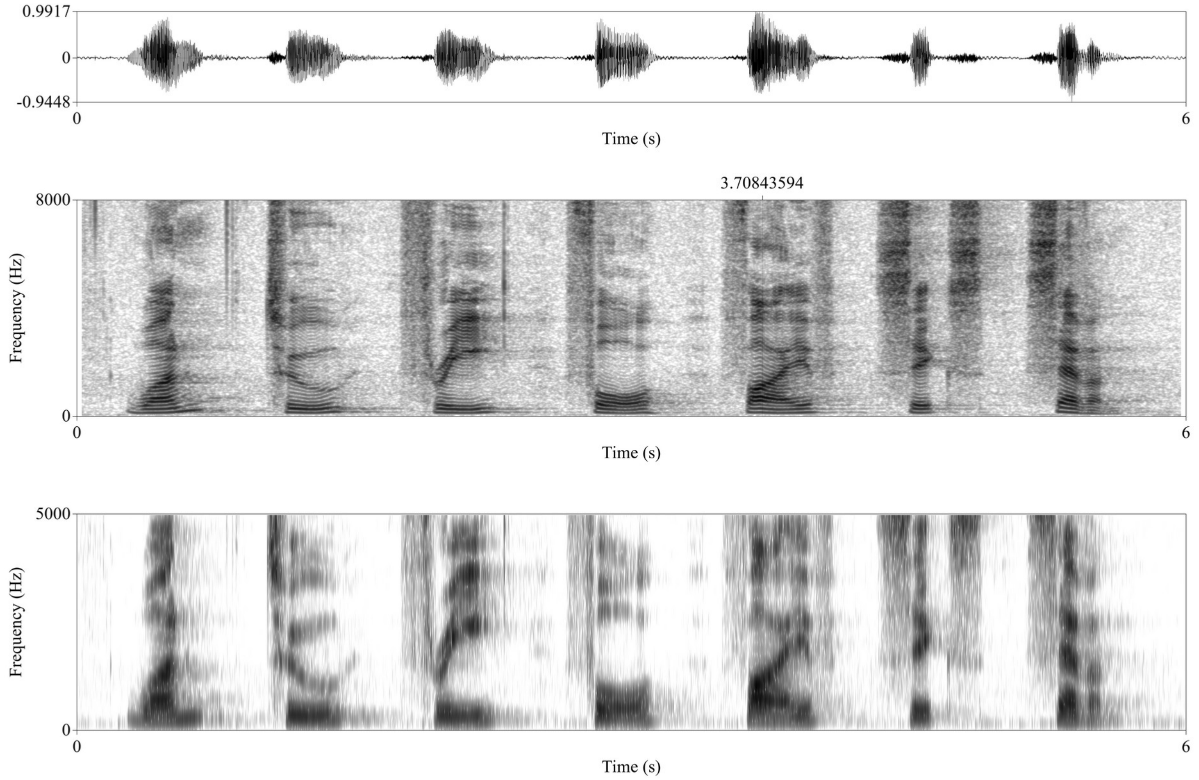
The amplitude of the harmonics varies at different frequencies in the spectrum, and a region of neighbouring stronger frequencies is termed *formant*. The formants result from the filtering function of resonances in the oral and nasal cavities, not unlike the equalizer filters of an audio amplifier system. These high frequency formants distinguish the timbre or perceived quality of different resonant sounds. In the example shown, the first [ɑ] formant is in a harmonic region around 650 Hz and the second is in*fundamental frequency* a harmonic region around 1.1 kHz, while the first [ɩ] formant is in a harmonic region around 550 Hz and 2 kHz. Formants can also be identified in unvoiced sounds, which have no fundamental frequency and therefore no harmonics, as regions of random noise frequencies instead of neighbouring regions of harmonics.

It is crucial to note that a *formant of a resonant sound* consists of a *neighbouring region of harmonics* of *F0*. The frequencies of *F0* and its harmonics, which relate to intonation, tones and pitch accents, are quite independent of the frequencies of the formants, which distinguish the qualities of voiced sounds. The essential difference between harmonics on the one hand and formants as regions of spectral frequencies on the other is sometimes not understood and is misrepresented in non-specialist accounts found on the internet, unfortunately even in a recently published textbook which focusses on phonology rather than phonetics.

A spectrum such as those in Figure 3 and Figure 4 is typically a ‘snapshot’ of a segment of the speech signal which has less than about 40 ms duration. Spectra can be calculated every 40 ms (or at other intervals) for the entire utterance and arranged in a series in order to represent the spectral changes throughout an utterance. For this purpose, the amplitude of the frequencies in the spectrum is shown in colour or on a grey-scale, with black representing the strongest and white representing the weakest frequencies.

In addition to the waveform, Figure 5 shows a combined *time domain* and *frequency domain* representation of the signal, the *spectrogram*, which consists of a sequence of spectra, each spectrum arranged with the time domain on the *y*-axis and the intensity represented on a grey-scale between black (most intense) and white (zero intensity). The broad dark lines represent the *formants* The correspondence between the waveform and the spectrogram on the time axis is evident, but the spectrogram shows far more detail than the waveform. It is evident that the [t] in *two* has high frequency aspiration, but no fundamental frequency, being unvoiced.

The middle graph is a *narrow band spectrogram*, which has high frequency resolution, showing both the harmonics as fine horizontal lines, and formants as darker regions of harmonics. The bottom graph shows a *broad band spectrogram*, which has a high temporal resolution and lower frequency resolution, showing only the formants and not the harmonics.



fun

Figure 5: Time domain and frequency domain visualisaitons of counting from one to seven in English (Praat Draw function): Top: waveform oscillogram, middle: narrow-band spectrogram, bottom: broad-band spectrogram.

damental frequency

2.5 Prosody

From a physical point of view, prosody (the tones, pitch accents, intonations, rhythms of languages) is represented by three parameters:

1. *timing* (measured in *seconds* or *milliseconds*, for example durations of syllables, words, phrases and longer units, and their rhythms),
2. *intensity* (amplitude squared, measured in *decibels* roughly corresponding to energy in production and perceived loudness),
3. *fundamental frequency* (measured in *hertz*, i.e. cycles per second, roughly corresponding to the rate of vocal cord vibration and to perceived pitch).

In some accounts of intonation, the term ‘prosody’ applies to segments of the speech signal which are at least the size of words, but in general, the term is applied to properties of speech segments which are at least the size of syllables. Speech segments which are shorter than syllables, i.e. consonants, vowels, glides and semivowels are usually dealt with separately.

The functions of prosody in communication are not the topic of the present contribution, but to put matters into perspective, the different kinds of function may be summarised in traditional linguistic terms:

1. Structural functions: intonation patterns mark the extents of grammatical units such as words and phrases, with delimitative boundary tones, configurative pitch contours and characteristic rhythms. In discourse structure, prosodic properties mark turn-taking and episode framing.
2. Semantic functions: pitch accents and tones mark contrasts between lexical items with high, low and contoured pitch patterns, and also with intensity variation between accented and unaccented syllables. Contrastive and emphatic pitch accents also mark discourse semantic properties such as informational focussing on new and given information in utterances.
3. Pragmatic functions: intensity, pitch and timing may also mark different kinds of speech act, and, consciously or subconsciously, they may also convey attitudinal and emotional meanings.

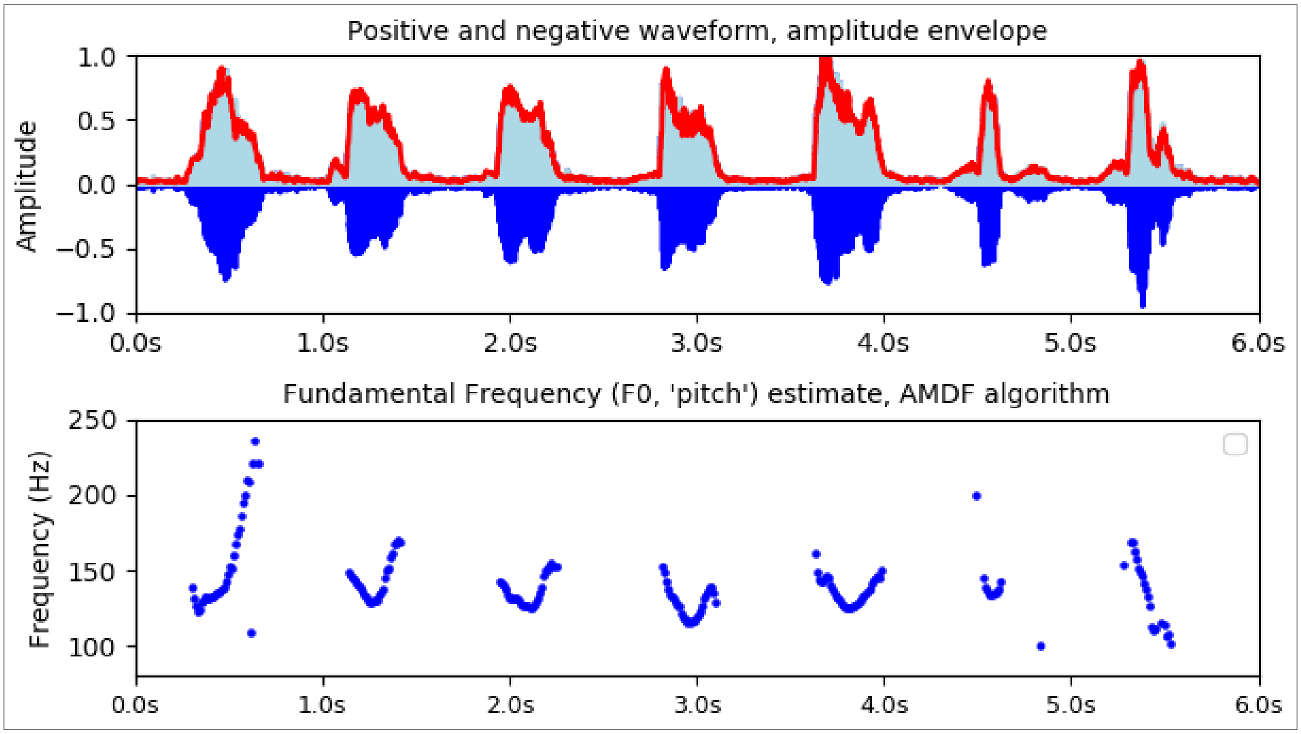


Figure 6: Time domain and frequency domain representations: top, waveform and amplitude modulation envelope; bottom, fundamental frequency estimation ('pitch track').

Figure 6 illustrates both time domain properties as well as basic frequency domain properties of speech, with the same information in the top graph and for the same recording as in Figure 2. The bottom graph shows the *fundamental frequency estimation*, a fairly complex procedure involving low pass filtering of the speech signal. Very short neighbouring segments of the speech signal (in this case only 10 ms long) are compared by subtracting their values in order to find out when the next most similar (strictly speaking: least different) short segment occurs. The function used in this case is the *average magnitude difference function* (AMDF). A similar function is used by the Praat software usingCheers correlation rather than subtraction, the *autocorrelation function* (ACF). One F0 calculation is performed every 10 ms for the visualisation in Figure 6, and representations of individual measurements are clearly visible in the figure as dots. Many algorithms have been developed for this purpose, in the time domain (like the AMDF and the ACF), or in the frequency domain by examining the frequency difference between harmonics in the spectrum.

Intonation is the perceived pitch pattern of an utterance from a grammatical perspective. Figure 6 shows a number of intonation features which are common in languages which do not have lexical tones or lexical pitch accents, such as English. In English the pitch accents associated with lexical stress positions may take on a variety of semantically and pragmatically determined pitch shapes which are organised as metalocutionary signs which denote the configuration of the grammatical unit (here simply a list of numerals) with which they are associated:

1. Globally falling contour, higher at the beginning than at the end.
2. Initial pitch accent with broad pitch range.
3. Final pitch accent with very low and narrower pitch level at the end (final lowering).
4. Iteration of the same pitch pattern among the medial pitch accents (traditionally the ‘head’ of the intonation pattern).

Figure 7 illustrates the very different kind of frequency distribution found in a lexical tone language, Beijing Mandarin, where the pitch pattern on a given word is lexically fixed and does not vary with grammatical, semantic or pragmatic functionality (though the global pitch contour may be affected by these factors). Figure 7 shows *F0* tracks for two speakers, each saying four monosyllabic morphemes: the syllable ‘ma’ and a different contrastive lexical tone, with the meanings ‘mother’ (tone 1), ‘hemp’ (tone 2), ‘horse’ (tone 3) and ‘scold’ (tone 4).

The visualisations in Figure 7 show that Speaker 1 has a higher-pitched voice (shown as higher *F0*) and greater pitch range (shown as the difference between lowest and highest *F0*). In addition, Speaker 2 has a different variety of tone 3, with *creaky voice* (also known as *vocal fry*) associated with no *F0* trace in the middle of the syllable. The creaky voice can be seen as an irregularity in the waveform of the utterance by Speaker 2. Creaky voice may have contrastive function in some languages, but in the case of Mandarin creaky voice is regularly associated with tone 3 for many speakers. Creaky voice is also frequently associated with low pitch in intonation patterns, and it may also occur as a general characteristic of a speaker’s voice, particularly with low-pitched female voices.

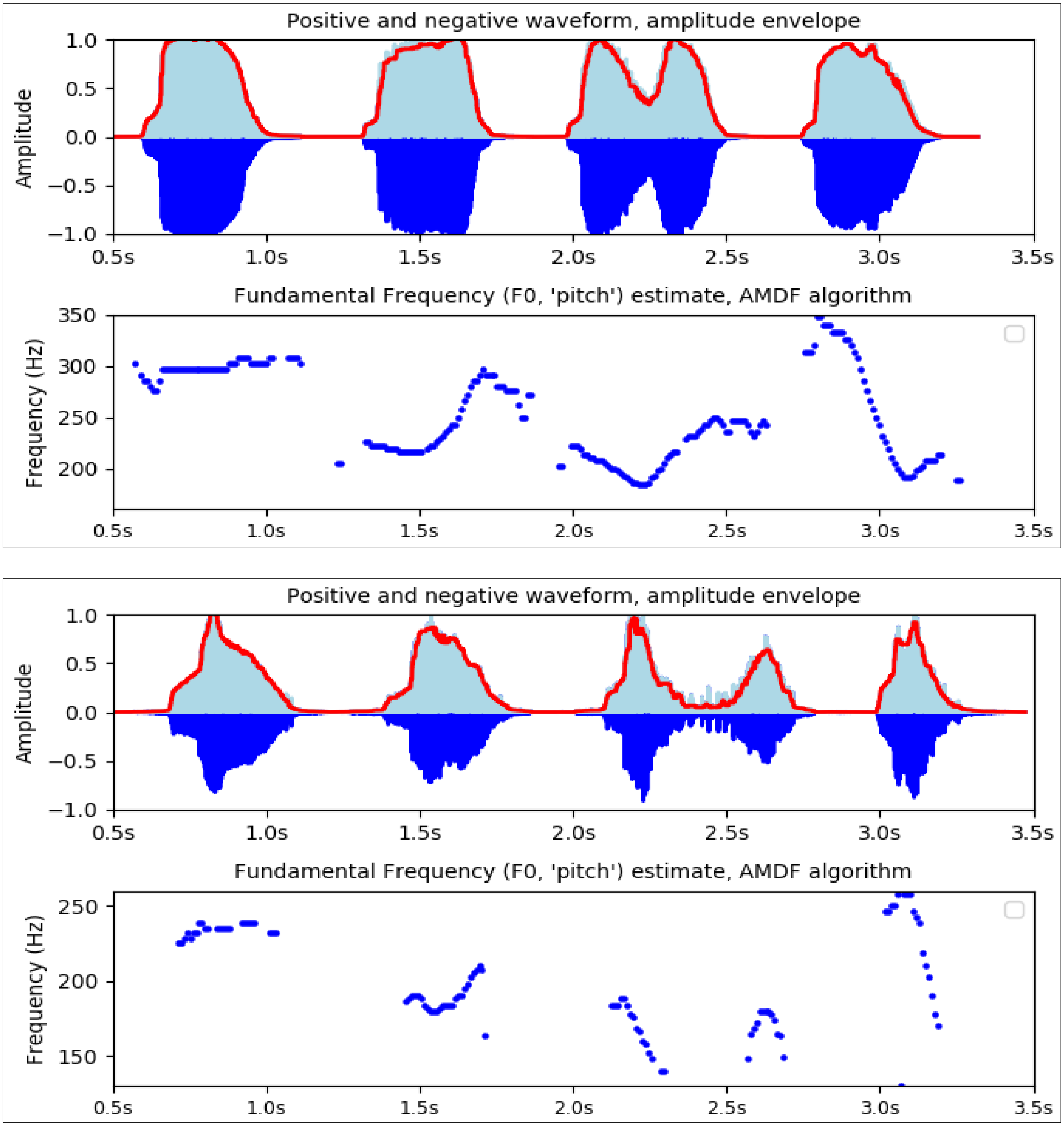


Figure 7: Beijing Mandarin, syllable 'ma', four tones: 1. high, 2. rising, 3. falling-rising (dipping), 4. falling.

Speaker 1 (top), female; tone3 with continuous F0.

Speaker 2 (bottom), female: tone 3 with gap in F0 aligned with creaky voice.

2.6 Voice quality, speech surrogates, gesture

Speakers of many, perhaps all languages use alternative acoustic and visual communication channels which all bear related functions to phonetic speech patterns and are commonly very relevant in fieldwork linguistics. They do not fall within the scope of the present contribution. Nevertheless, the main features are outlined here in order to provide a broader semiotic context for phonetic fieldwork.

Voice quality (informally characterised as ‘gentle’, ‘harsh’, etc.) may be an indexical characteristic of an individual speaker, or an expression of emotionality or situational context. Breathy voice, which is also used as a lexical phonetic feature in some languages, may be associated with excitement of various kinds, or with a ‘panting’ variety of speech after physical exertion.

Conversational gesture, also known as co-speech gesture or gesticulation, accompanies speech, which is of course also gesture of the vocal organs. Conversational gesture adds to or emphasises (and sometimes contradicts) the content of the speech channels in different ways:

1. iconic gestures demonstrating physical shapes and sizes of the item referred to (cf. onomatopoeic sounds in speech),
2. deictic gestures point to the position of the item referred to (cf. phrasal accents in speech),
3. beat gestures demonstrate rhythms or regular emphases (cf. speech rhythms of syllables or words),
4. symbol gestures demonstrate specific objects or actions such as waving goodbye or particular finger actions for agreement or good luck (cf. the conventional lexicon of spoken language).

Each of these gestures may have metaphorical uses, such as the ‘size’ of an emotional reaction. Some conversational gestures are accompanied by acoustic effects, as in lip-smacking and tongue-clicking, clapping, finger-snapping and stamping. Each of these gestures may have metaphorical uses, such as the ‘size’ of an emotional reaction. The temporal alignment of conversational gesture and speech is significant, with a gesture often starting before the associated word or syllable. The semantic relation between conversational gestures and speech sounds may be rather close: accented syllables may be accompanied by movements of the hands, or by eyebrow-raising, for example, and an ejaculative expression such as “Oooh!” involves visible lip-rounding.

Similar conversational gestures have very different meanings in different cultural communities. For example, forming a ring with the tips of the thumb and the first finger may signify success in some communities, while in other communities it may be an obscene insult. The same applies to many other gestures, and in fieldwork studies the possibility of gestural taboos should not be ignored.

Speech surrogates are situationally dependent alternatives to speech communication, such as whistling to draw attention. Some speech surrogates form complex systems, such as whistling communication based on the tones and intonations of a language, or drumming, also based on the tones and intonations of a language. These speech surrogate channels are often associated with restricted registers of language use such as religious and other cultural ceremonies. Care is needed in dealing with speech surrogates. It is not well-known, for example, that in some societies whistling is regarded as indecent or even as a tabooed obscenity. This may limit elicitation choices when investigating tones, for example.

Signing, or sign language, as a communicative medium among the auditorily impaired is fundamentally different from conversational gesture, though it may also co-occur with conversational gesture, like speech. Signing is a form of language with its own complex grammatical structure and complex lexical gestures, and there is a large body of literature concerned with the different sign languages in different communities all over the world. Sign languages are not universal, though they may be based on universal principles. Even in a single cultural area there may be locally different sign language dialects or even different sign languages.

3 Data gathering: systematic planning, recording

3.1 Useful tools

There are many software tools which are useful for gathering and processing fieldwork data. For audio recording, a good pocket-sized digital audio recorder is often used, with the quality needed for music recording; dictation quality hardware is not easily usable and often uses unusual audio formats. For recording video, in addition to recording with a video camera (sometimes a smartphone is sufficient), parallel recording with an audio recorder can be very useful. For both audio and video recording it is sometimes adequate to use a laptop with appropriate software. A small selection of specific useful software tools are listed here; tools for other aspects of linguistic fieldwork such as morphological and syntactic text annotation, lexicon construction, concordancing and database management, are discussed in other contributions.

1. Praat: the Praat software tool is interoperable on current PC platforms (Windows, Linux, MacOS), and has become a standard tool for recording, analysis and synthesis of the acoustic properties of speech (and other acoustic data), including fundamental frequency estimation (‘pitch extraction’) and intensity visualisation. Praat also has its own programming language (‘Praat scripting’) for automatising many kinds of analysis and a function for generating publishable diagrams. For this reason Praat is often referred to as a ‘phonetic workbench’.[1]
2. ELAN: For the study of the situative context of speech utterances, and of gesture and signing, the ELAN multimodal annotation tool for video and audio recordings is frequently used in linguistics, phonetics and gesture studies. ELAN can export audio to Praat for more detailed analysis. ELAN is also interoperable on Windows, Linux and MacOS.[2]

For each of these tools the associated web sites provide introductory and user guide materials. There are many sites for Praat scripts which can be found by searching the internet. Both Praat and ELAN are designed for analysis of the transmission phase of speech, not directly of speech production

[1] https://www.fon.hum.uva.nl/praat/

[2] https://archive.mpi.nl/tla/elan

and perception. For the analysis of speech production in the field, numerous tools are available, ranging from the laryngograph or glottograph for measuring events in the larynx, as well as instruments for measuring airflow and tongue position. These methods are outside the scope of the present contribution; this applies also, evidently, to X-rays and invasive methods which require medical expertise and specific ethical approval.

For speech data elicitation and a first impressionistic analysis, there are many informal tips and tricks which are used by linguists and phoneticians in order to provide informal acoustic models of lexical and phrasal prosody for the listener. For example, humming, whistling and musical instruments such as slide flutes, violins or kazoos, and to a limited extent lamellophones (‘jaws harps’) can be used to ‘model’ tones and intonation (bearing in mind that there may be social taboos on whistling and music, depending on local cultural and religious conventions). Lightly touching the larynx with a finger may be helpful in identifying tonal movements (again bearing touch taboos in mind). Holding a piece of light-weight paper in front of the lips or the nose can also be helpful in confirming aspiration of stops. Asking the speaker to describe their own impressions of the sounds can be very helpful (though scientifically not too informative), for example expressions such as ‘sticky sounds’ for implosive consonants, or ‘sing-song’ for tones, pitch accents and intonation.

3.2 Pre-recording phase: data planning

The many different strategies of data acquisition fall into three broad categories: participant observation with or without recording of participant interactions, and directed observation with questionnaires or strict experimental paradigms, evaluation and deployment of legacy materials.

The first decision to make is which acoustic features of the speech signal are important (e.g. pitch, formants). For example, there are different linguistic and phonetic definitions of 'stress'. Stress is not a straightforward acoustic phonetic property but a complex category which may be phonological, or a factor in speech production, or a prominently perceived item in an utterance. Some factors are valid for all purposes and include many of the following:

1. Speech styles and registers to be recorded:
    1. Prompted ‘laboratory’ speech for phonetic experiments.
    2. Task-oriented speech.
    3. Dialogue (interview, conversation, etc.): it is important to note that dialogues for phonetic analysis require stereo recording.
2. Technical choices:
    1. Microphone (to capture the range of frequencies required).

2. Choice of audio file format (e.g. WAV for full quality recording, vs. MP3 or WMA compressed formats for lower quality but smaller files); however, in general WAV format should be used for phonetic analysis with Praat.
3. Choice of recorder with choice of WAV and MP3 formats (e.g. Zoom H2N Handy Recorder, the most popular in general, or Roland R-05, popular among musicians); a recorder with only MP3 or WMA formats may be suitable for some purposes, and for the analysis of stress it may be suitable when looking at syllable duration and fundamental frequency.
4. Choice of recorder setting:
    1. Volume:
        1. no Automatic Gain Control (AGC): if you use it, you will not be able to use measurements of amplitude or intensity since AGC changes these automatically to keep the volume constant (also resulting in changes of noise level),
        2. set the volume control as high as possible, but avoiding peak overdrive (i.e. the peaks should be just below the maximum or ‘red’ level on the volume meter).
    2. Sampling frequency:
        1. sampling frequency: at least 2*f*, where *f* is the highest frequency to be recorded (Nyquist sampling theorem),
        2. high quality: human hearing reaches about 20kHz, so high quality recordings must use at least 40kHz sampling frequency (e.g. the 48kHz DAT standard or the 44.1kHz CD standard),
        3. medium quality: 40kHz is not strictly necessary for most phonetic purposes, as the relevant frequencies (fundamental frequency below about 600Hz (child speech), vowel formant frequencies below 6kHz) are generally below 10kHz, so 20kHz is sufficient, and in practice 16kHz and 22.05kHz are often used,
        4. oversampling: oversampling uses a sampling rate which is much higher than the Nyquist frequency in order to avoid various noise effects; for phonetics, the best compromise frequency is a Nyquist frequency of 40kHz, with the standard sampling frequencies of 48kHz or 44.1kHz,
        5. sampling curiosity: in case you are curious about why the strange-looking number 44.1kHz was chosen as a standard: it is the product of the squares of the first 4 prime numbers above 1, which permit efficient ‘down-sampling’ to 16 different lower frequencies, basically by dividing by combinations of these numbers:

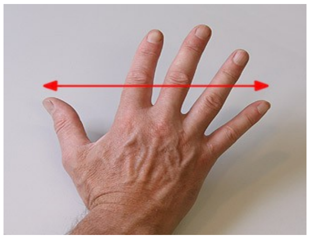
            2^2 x 3^2 x 5^2 x 7^2 = 44100Hz

3.3 Recording phase

There are several points to be considered during the recording phase, in particular the scenario environment, which will be depend on the speech style being recorded:

1. *Scenario environment for recording*:
    1. Avoid echo as far as possible:
        1. echo is caused by hard walls and floors, so an environment with soft furnishings (curtains, carpets, cushions) is preferable, if recording is done outside a studio;
        2. a studio should have sound-proofed floor, walls and ceiling.
    2. *Avoid noise*:
        - Place the microphone at least 25cm (10”) from the speaker in order to avoid breath noise, if possible slightly to the left or the right of the speaker,

          A useful measure for the minimum distance between microphone and mouth is the span, the distance between outstretched thumb and little finger, which is usually between about 18cm and 22cm for adults, and only a little less than 25cm.

          
        - When recording outside use a sheltered place to avoid wind and other noises; note that a wind muff (wind shield, ‘dead cat’) may filter out high frequencies, though this may not be important for some purposes (e.g. for news reporting).
2. Speaker:
    1. The speaker should be asked to give permission for the recording to be used for scientific purposes.
    2. The permission can be in writing, but in any case should be included in the recording as recorded spoken metadata.
    3. The recording should include other metadata, including the date, the place, the speaker(s), other participants such as audience and those making the recording.
    4. The speaker should receive appropriate instructions.
    5. Speakers should take a sip of water every 5-10 minutes to avoid drying out the vocal folds and thereby changing the voice quality.
3. Protocol:

1. A protocol of the recording session should be kept, containing:
    1. the same metadata information as on the recording (time, place, participants),
    2. instructions given to speaker(s),
    3. permissions of speaker(s)
    4. file names and any other aspects of the recordings,
    5. a list of any problems which occurred.

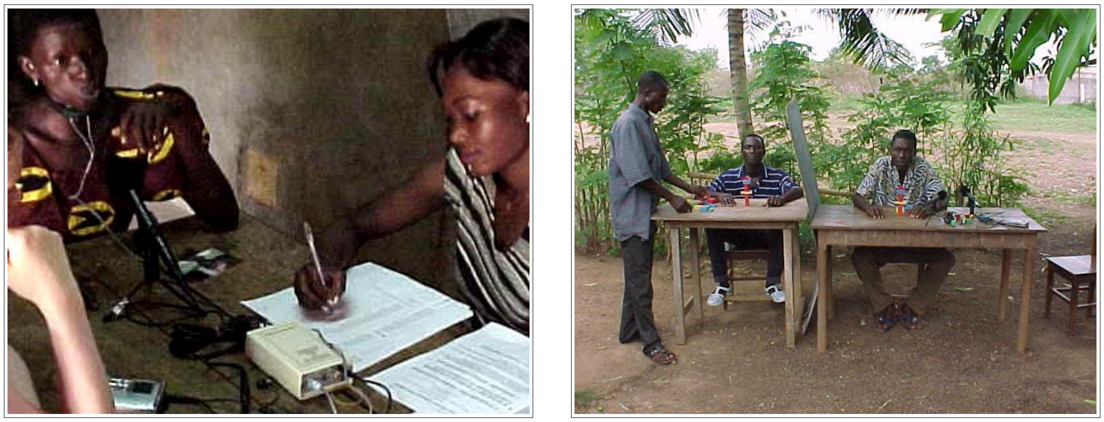


Figure 8: Two phonetic fieldwork scenarios:
Left: recording session with digital recorder, laryngograph and wordlist.
Right: Dialogue recording of blocks world negotiation.

3.4 Post-recording phase

There are many kinds of post-recording activity, including the following:

1. Systematic filing of metadata in separate documents, preferably in a database (or temporarily in a word processor table).
2. Systematic labelling of recorded audio files with a project (or language, etc.) name, serial number, and date, e.g.: ‘englinterview_05_2016-05-09.wav’. Spaces should not be used in filenames; the underscore (“_”) should be used instead.
3. Cutting of recordings: usually necessary in order to systematise items for analysis – but always keep the original recordings (cutting can be done using the general audio tool Audacity or the phonetic workbench Praat).

The first step after ensuring sustainable storage of the data is the documentation of metadata (recording filename, date, time, place, participants, recording equipment, recording situation). The metadata filename should also reflect the filename of the recording. There are recommended formats for metadata, the most well-known in linguistics and phonetics being the format of the Open Language

Archives Community (OLAC),[3] which also provides a portal for registering linguistic and phonetic metadata in order to support communication and interchange in the scientific community.

The next step in the post-processing of recordings is transcription, the representation of speech events in writing, and annotation, the enhancement of transcriptions with time-stamps indicating the positions of speeech events in the recording. Except for very approximate purposes, linguistic and phonetic annotation require the use of specialised software such as Praat, AnnotationPro,[4] WaveSurfer or ELAN, and semi-automatic transcription support with SPPAS.[5] Transcription types vary according to needs (e.g. IPA or SAMPA phonetic vs. orthographic vs. discourse analytic transcription), labelled to match the recording. It is often convenient to perform transcription and annotation simultaneously. Annotation not only provides a very useful starting point for locating items in the data, the time-stamp information is an essential source of information for investingating the durations of speech sounds and larger units of speech, of duration regularity and of speech rhythm. This transcription and time-stamp information can be extracted, for instance from a Praat annotation file, with a fairly straightforward program, for instance with the programming language Python, in order to perform further analysis of the transcription text or, using a spreadsheet application such as Excel or LibreOffice Calc, or statistical software, for further phonetic analysis.

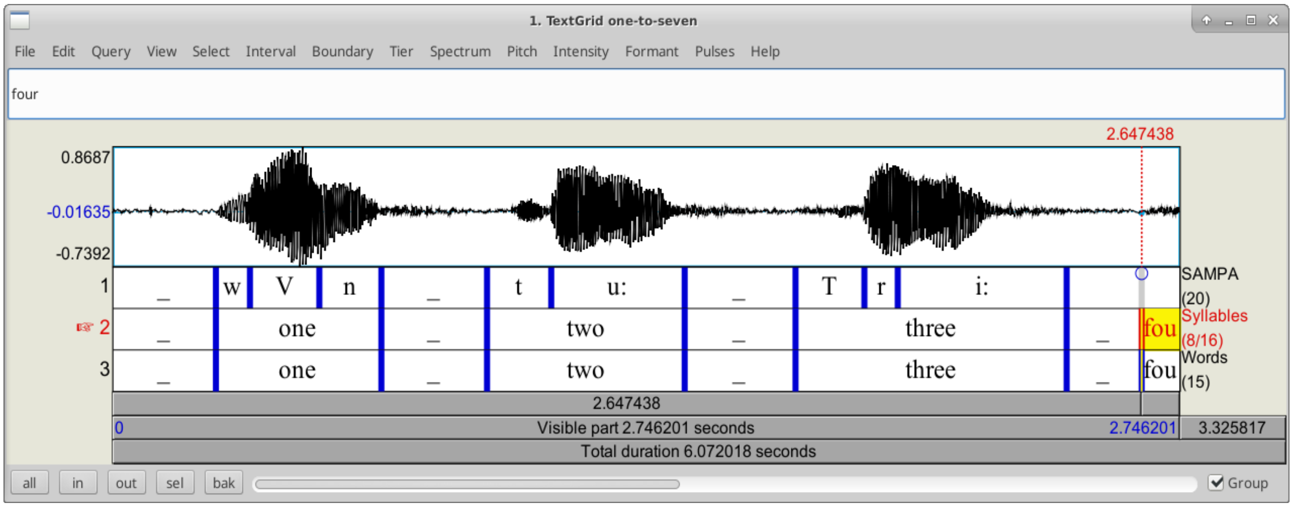


Figure 9: Interactive annotation with Praat graphical user interface.

Figure 9 shows interactive annotation with the Praat graphical user interface in progress. The waveform oscillogram is shown at the top, and annotation tiers of different types are aligned below the waveform. In Figure 9 the tiers (from the bottom): Words, Syllables and SAMPA. Clicking on the little circle on the vertical cursor line inserts a new boundary.

SAMPA is relatively often used for easy entry of IPA transcriptions. SAMPA stands for *Speech Assessment Methods Phonetic Alphabet*, and was designed in a European speech technology project to

[3] OLAC: http://www.language-archives.org/

[4] http://annotationpro.org/

[5] http://www.sppas.org/

allow a keyboard-friendly easy and rapid entry and machine processing of IPA transcriptions. There is a one-to-one mapping between SAMPA characters, and IPA characters, so SAMPA is simply an alternative set of glyph shapes for the IPA, not a different alphabet. Later an extension, X-SAMPA, was developed by Wells to cover the entire IPA character set (cf. Gibbon et al. 1997; Gibbon et al. 2000 for complete specifications). There are several SAMPA-IPA converters on the internet. Whether the standard IPA glyph shapes or SAMPA are used is a matter of convenience, preference and intended application.

It is essential to keep a written record of the metadata, the transcription and the annotation. For this purpose, Unicode UTF-8 is adequate for many languages; for some scripts, the more complex Unicode encodings are needed.

3.5 Phonetic analysis

A standard reference textbook for experimental and instrumental phonetic data analysis in the fieldwork contexts is by Ladefoged (2003). The equipment mentioned (e.g. tape recorders) is a little outdated, but the methods and topics have otherwise not changed. The book covers …

The phonetic workbench Praat has many uses in many disciplines, ranging from the study of specific speech sound properties such as voicing, aspiration, nasality, vowel type to the study of lexical tones and pitch accents, and of phrasal intonations. In addition to the initial steps of recording, transcription and annotation, Praat provides a wide range of methods, from the simplest to the most sophisticated, for scientifically analysing the phonetic properties of recorded speech. A wide selection of useful Praat scripts (programs) for automatic extraction of phonetic properties from annotated speech recordings have been made available by many phoneticians and can easily be found on the internet. A wide range of easily accessible phonetic tools have also been made available on the internet, including several by the author of this contribution (e.g. Time Group Analyzer; cf. Gibbon and Yu 2015).

4 Summary, conclusion and outlook

The context of phonetic fieldwork is varied and may be addresses with varying degrees of complexity, depending on the goals of the fieldwork, ranging from orthography plus prosodic markings for interpretative studies in sociolinguistics and conversational analysis to studies to detailed experimental acoustic studies using advanced statistical methods. The aim followed in the present contribution is more modest, concentrating on basics of phonetics and first stages of phonetic data collection and processing with fieldwork and the first stages of data analysis in the field in mind. The aspects covered include a general outline of phonetic concepts concerning speech sounds and prosody, with notes on related topics such as speech surrogates and gestural communication, as well as basics of sound recording in the field for phonetic purposes.

Linguistic and phonetic fieldwork are skill sets with many dimensions, ranging from intellectual and technical preparation through obtaining permissions, negotiation with the language community and language consultants to the special skills required in eliciting data and operating equipment and then the linguistic and phonetic processing of the data, analysis, theoretical description and explanation, publishing and return of results to the language community. The present contribution outlines some of the specifically phonetic skills which are needed.

5 References

Gibbon, Dafydd and Jue Yu. 2015. *Time Group Analyzer: Methodology And Implementation*.The Phonetician 111/112:9-34.

Gibbon, Dafydd, Roger Moore and Richard Winski, eds. (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Gibbon, Dafydd, Inge Mertins and Roger Moore, eds. (2000). *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Dordrecht: Kluwer Academic Publishers.

Hayward, Katrina. 2013. *Experimental Phonetics: An Introduction*. London: Routledge.

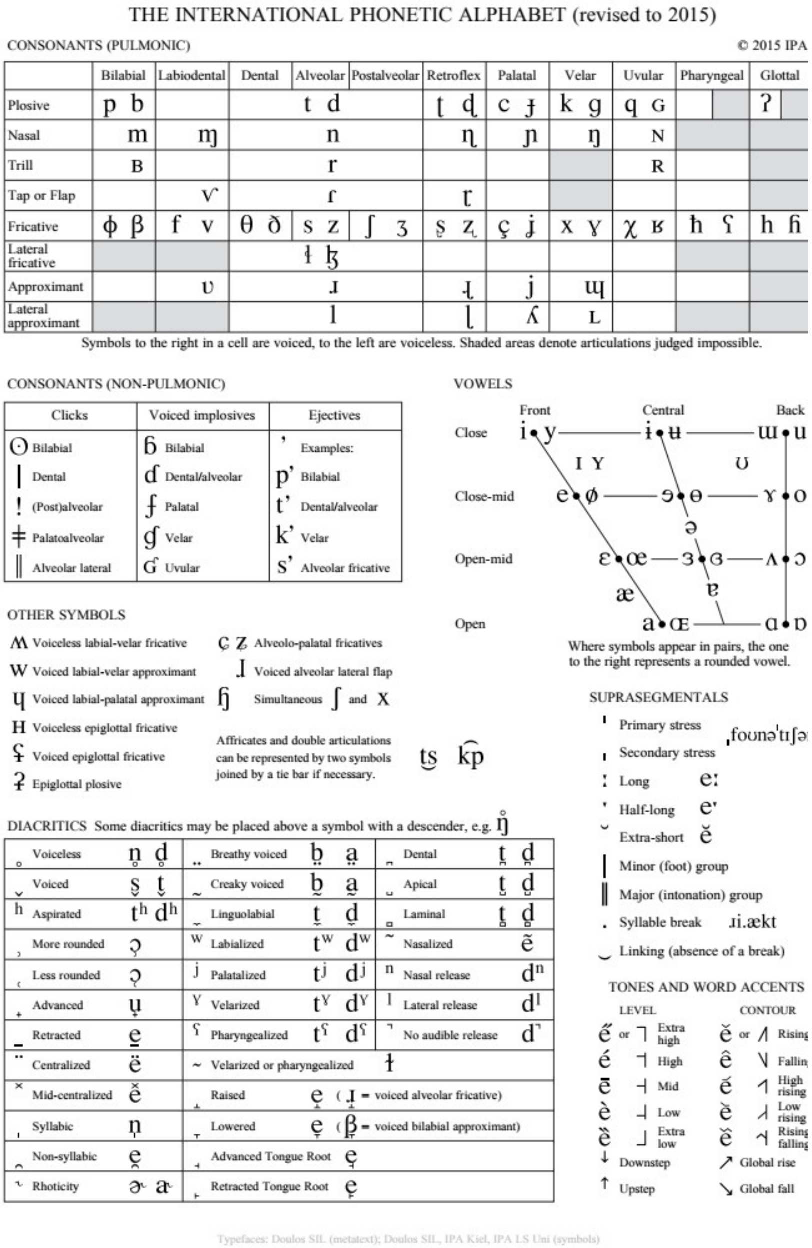
International Phonetic Association. 1999. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge, U.K.: Cambridge University Press. https://www.internationalphoneticassociation.org/content/full-ipa-chart

Ladefoged, Peter. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Malden, MA, & Oxford: Blackwell, 2003.

McKinney, Norris P. and Carol V. McKinney. 2017. *An Introduction to Field Phonetics*. SIL International Publications.

Reetz, Henning and Allard Jongman. 2020. Phonetics: *Transcription, Production, Acoustics, and Perception*. Blackwell Textbooks in Linguistics Book 35. London: Blackwell. Second edition.

6 Appendix: International Phonetic Alphabet



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	ʼ	Examples:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

VOWELS

	Front		Central		Back
Close	i • y		ɨ • ʉ		ɯ • u
		ɪ ʏ		ʊ	
Close-mid	e • ø		ɘ • ɵ		ɤ • o
			ə		
Open-mid	ɛ • œ		ɜ • ɞ		ʌ • ɔ
		æ	ɐ		
Open		a • ɶ			ɑ • ɒ

Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ɧ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. t͡s k͡p

SUPRASEGMENTALS

ˈ	Primary stress	ˌfoʊnəˈtɪʃən
ˌ	Secondary stress	
ː	Long	eː
ˑ	Half-long	eˑ
˘	Extra-short	ĕ
|	Minor (foot) group	
‖	Major (intonation) group	
.	Syllable break	ɹi.ækt
‿	Linking (absence of a break)	

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	~	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝ (ɹ̝ = voiced alveolar fricative)			
◌̩	Syllabic	n̩	◌̞	Lowered	e̞ (β̞ = voiced bilabial approximant)			
◌̯	Non-syllabic	e̯	◌̘	Advanced Tongue Root	e̘			
˞	Rhoticity	ɚ a˞	◌̙	Retracted Tongue Root	e̙			

TONES AND WORD ACCENTS

LEVEL				CONTOUR			
e̋	or	˥	Extra high	ě	or	˩˥	Rising
é		˦	High	ê		˥˩	Falling
ē		˧	Mid	e᷄		˦˥	High rising
è		˨	Low	e᷅		˩˨	Low rising
ȅ		˩	Extra low	e᷈		˧˦˧	Rising falling
↓	Downstep			↗	Global rise		
↑	Upstep			↘	Global fall		

Typefaces: Doulos SIL (metatext); Doulos SIL, IPA Kiel, IPA LS Uni (symbols)

Figure 10: International Phonetic Alphabet (revised to 2005).
Cf. International Phonetic Association (1999).