Phonetic Transcription and the International Phonetic Alphabet

Claire Brierley and Barry Heselwood, University of Leeds

https://doi.org/10.1093/acrefore/9780199384655.013.892
Published online: 23 March 2022

Summary

Phonetic transcription represents the phonetic properties of an actual or potential utterance in a written form. Firstly, it is necessary to have an understanding of what the phonetic properties of speech are. It is the role of phonetic theory to provide that understanding by constructing a set of categories that can account for the phonetic structure of speech at both the segmental and suprasegmental levels; how far it does so is a measure of its adequacy as a theory. Secondly, a set of symbols is needed that stand for these categories. Also required is a set of conventions that tell the reader what the symbols stand for. A phonetic transcription, then, can be said to represent a piece of speech in terms of the categories denoted by the symbols. Machine-readable phonetic and prosodic notation systems can be implemented in electronic speech corpora, where multiple linguistic information tiers, such as text and phonetic transcriptions, are mapped to the speech signal. Such corpora are essential resources for automated speech recognition and speech synthesis.

Keywords: phonetic notation, phonetic transcription, International Phonetic Alphabet, speech recognition and synthesis, automatic alignment and annotation

Subjects: Phonetics/Phonology

1. Introduction

As a form of writing, phonetic transcription is *technographic*, not orthographic (Crystal, 1987, p. 194). The function of an orthography is to provide spellings enabling the identification of language-specific lexical items in written texts. The aim of a phonetic transcription, however, is not to identify lexical items in spoken texts but to express an analysis of the phonetic structure of the speaker's utterance in terms of the categories of a universal phonetic theory. It can therefore claim to be as scientific as other technographic writing systems such as algebraic or chemical notation. In the words of David Abercrombie, a 20th-century authority on phonetics and phonetic transcription, "phonetic transcription records not an utterance, but an analysis of an utterance" (Abercrombie, 1967, p. 127), that is to say, an analysis in terms of the categories of phonetic science. To mark a transcription as phonetic, it is enclosed in square brackets [. . .].

The phonetically rich and intrinsically dynamic nature of speech cannot be exhaustively captured in a transcription. Phonetic transcription is therefore a process of data reduction. However, it can also be said that a phonetic transcription, because it expresses an analysis, contains more than is contained in the speech signal. By way of example, the transcription [bʊk] provides an analysis of a pronunciation of the word *book* into the phonetic category-bundles 'voiced bilabial plosive', 'mid-close back rounded vowel', and 'voiceless velar plosive'; needless to say, the speech signal carrying that pronunciation contains no such analysis.

2. Notation Systems

Phonetic notation has its origins in the development of *phonographic writing*. In phonographic writing the elements of the writing system correspond to elements of pronunciation such as syllables, consonants, and vowels, so that it takes only a change of perspective to regard written forms as representing the corresponding spoken forms in addition to identifying them as lexical items. Logographic writing does not provide for this possibility; there is no indication in the Chinese character 书 'book', for example, that its pronunciation is [ʂu˥] (Pinyin 'shū').

Alphabets are well suited to being adapted for phonetic notation, having arisen in the first place through attention to pronunciation via rebus writing, syllabic writing, and acrophony. The more regular the sound–spelling correspondences in a language's orthography, the more it is possible for alphabetic letters to function as a notation for consonants and vowels.

Although the ancient Indians and Greeks, and the grammarians of the Abbasid period in the medieval Middle East, achieved much in understanding the phonetics of speech, they were not concerned with developing phonetic notation much beyond what their orthographies provided. The development of phonetic notation as it is currently known can be traced back to Western Europe in the early modern period. The story begins with the adoption, diffusion, and adaptation of the Roman alphabet among speakers of languages in which sound–spelling correspondences were, or became over time, more irregular and more complicated than those of Latin, such that it was not possible to represent pronunciation satisfactorily using only the letters bequeathed to them.

An early and isolated example of a comprehensive vowel notation system constructed on a Roman alphabet base is that of the anonymous 12th-century Icelandic 'First Grammarian' who employed diacritics for nasalization, length, and vocalic openness (Haugen, 1972). It was not, however, until the 16th and 17th centuries that a concerted attempt to develop resources for representing pronunciation arose among members of the "English School of Phonetics" (Firth, 1946). The Roman alphabet had never fitted English phonology particularly well, but the mismatch had become more acute in early modern English than in most other European vernaculars due to the Great Vowel Shift, which had begun in the early 1400s and was more or less complete by the late 1600s. Motivated by spelling reform, and later by the project for a universal language (Firth, 1946), these scholars experimented with new letter designs to try to achieve spellings that adhered to a 'one letter, one sound' principle (Salmon, 1972). Much progress was also made in understanding the phonetics of speech, which led to descriptions of letters in phonetic terms, in effect transforming them into phonetic symbols. In their charts, letter-symbols can be seen as products of the intersection of categories, which is how symbols are defined on modern International Phonetic Association (IPA) charts (IPA, 1999, p. 159). By the close of the 17th century, then, the possibility of defining phonetic symbols in terms of theoretical categories had been established, and the principle of 'one sound, one symbol' clearly articulated.

The 18th century saw comparatively little by way of advances in phonetic theory and notation, but significant strides were made in the second half of the 19th century, culminating in the formation of the IPA and the first version of its alphabet chart. In Section 2.1, principles of design that have influenced notation systems are discussed, and examples of their presence in IPA notation are identified.

2.1 Design Principles of Notation Systems

Notation systems have been constructed according to various principles. Because more than one design principle can be seen in most notation systems, it is perhaps more insightful to talk of these than to try to typologize notations too strictly. In discussing and illustrating these principles, the focus will be on IPA notation, but their precursors in earlier notation systems will also be identified.

2.1.1 Organicity

The principle of *organicity* in phonetic notation means that phonetic symbols represent those speech organs thought to be responsible for the production of the sounds being symbolized, and the positions they take up relative to each other. Sweet (1881, p. 177) cited Alexander Melville Bell's *Visible Speech* (1867) notation as an organic alphabet, and Sweet himself went on to develop his own version of one (Sweet, 1906), as did Daniel Jones in collaboration with Paul Passy, a major figure in the formation of the IPA (Passy, 1907). The symbol [

] in Bell's notation is equivalent to IPA [d]. The [

] part stands for 'point of tongue contracting mouth-passage' (Bell, 1867, p. 39), while the bar across the top shows that the mouth-passage is closed, and the vertical line represents the vocal folds in the voiced state (Bell, 1867, pp. 38–39). Further examples of Bell's notation are given and discussed in Sections 2.1.2 and 2.1.3.

In the IPA, organicity is evident in the conventions rather than the notation, in that symbols are defined largely in terms of places and manners of articulation, relying heavily on anatomical terminology.

2.1.2 Analogy

The principle of *analogy* is the principle that a given symbol-component should always denote the same phonetic category, and that a given phonetic category should always be denoted by the same symbol-component. It thus takes the ‘one sound, one symbol’ principle down to the level of the phonetic feature. A well-known historical example of an analogical notation is that designed by Bishop John Wilkins (Wilkins, 1668, p. 376). His chart, reproduced in Figure 1, presents symbols constructed according to relationships of place and manner of articulation. Plosive symbols, for example, all comprise a vertical line to which a line is adjoined at the top left for voiced and at the base right for voiceless: angled for bilabial ([ᒣ] = [p], [ᒣ] = [b]), horizontal for alveolar ([ᒣ] = [d]), and crossed for velar ([T] = [g]).

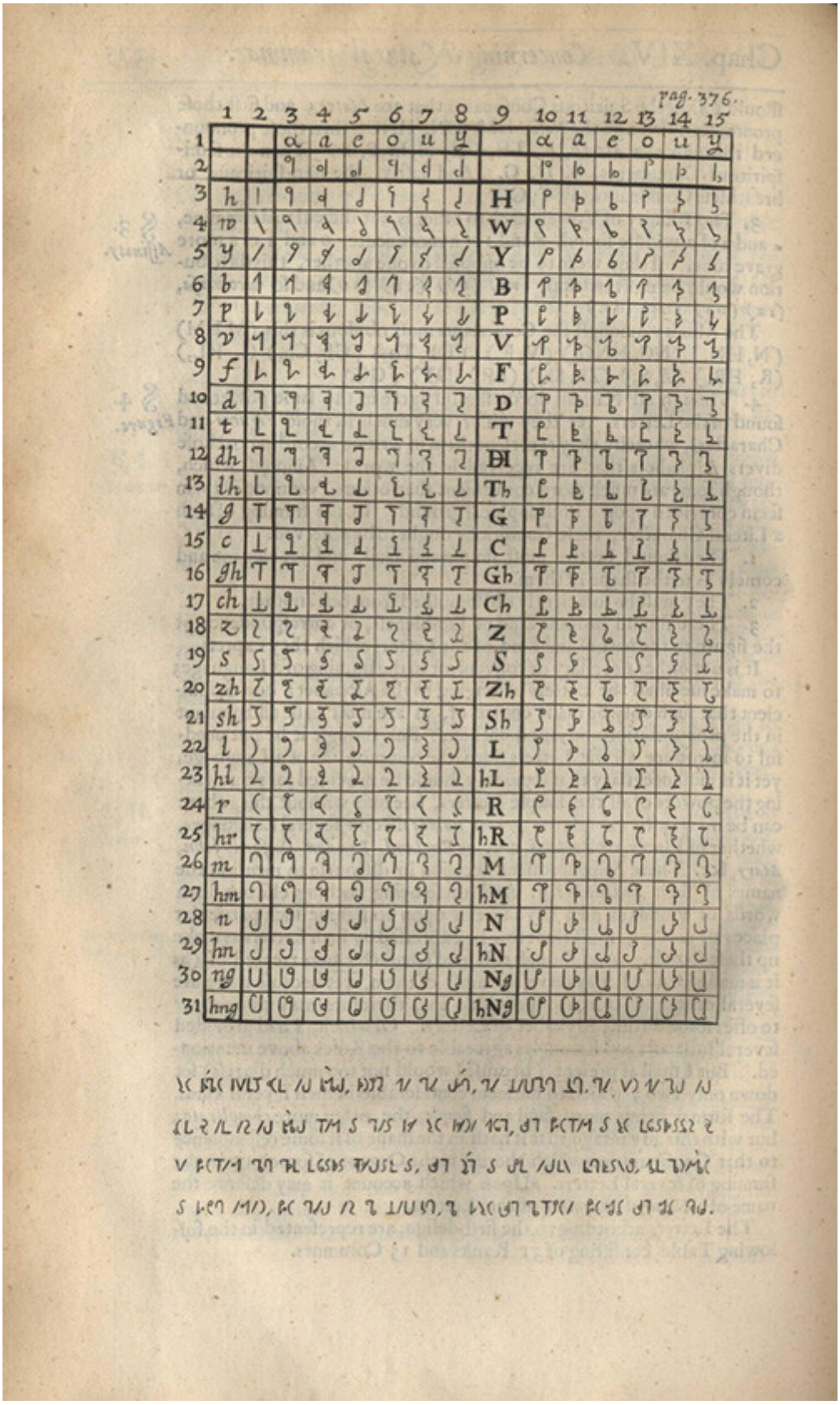

Figure 1. Wilkins's symbol chart for consonants and vowels (Wilkins, 1668, p. 376).

Source: Reproduced with the permission of Special Collections, Leeds University Library *<https://explore.library.leeds.ac.uk/special-collections-explore/213010>*

In Bell's *Visible Speech* notation, analogy is evident throughout the construction of complex symbols. For example, [Ɑ] represents a voiceless back stop (= IPA [k]) and [

] a voiced back stop ([g]); [Ɒ] represents a voiceless labial stop ([p]), and [

] a voiced labial stop ([b]). These symbols are constructed systematically from the components [C, Ɔ, I, **I**], which denote, respectively, 'back consonant', 'labial consonant', 'shut' (i.e., a stop), and 'voice' (the latter oriented horizontally).
The analogical principle is evident in the IPA, for example in the symbols for retroflex consonants, which have in common a descending right tail, in the use of an apostrophe to denote ejectives, and of an ascending right hook for implosives. Analogy tends to be behind the modification of Roman letters to form new IPA symbols. The voiced palatal consonant symbols derive from <j> with a descending left hook, for example, [ɟ] for 'voiced palatal plosive'; rhotic symbols from <r, R>, for example, [ɹ] for 'alveolar approximant'; laterals from <l, L>, for example, [ɬ] for 'voiceless alveolar lateral fricative'; and nasals from <m, n, N>, for example, [ŋ] for 'velar nasal'.

Analogy is probably at its most explicit in the use of *diacritics*. Diacritics have often been a source of controversy in phonetic notation, but the alternative for a notation system based on the letters of the Roman alphabet is either to proliferate digraphs and other multiletter symbols or to recruit a substantial number of new and unfamiliar symbols (Albright, 1958). In the 19th century, Richard Lepsius chose to use diacritics extensively in his *Standard Alphabet* (Lepsius, 1863), designed for universal application. By contrast, his contemporary A. J. Ellis avoided diacritics in favor of multi-letter symbols and a mixture of Roman and italic faces in his palaeotype notation (Ellis, 1867). Analogy can never be as transparent in multi-letter symbols as it can in base symbols modified systematically by diacritics, but there is often a certain arbitrariness in whether a phonetic feature is denoted by a diacritic or a base symbol. In the IPA, for example, dental articulation is denoted integrally in [θ ð] but by diacritics in the dental stops [t̪ d̪] and interdental stops [t̪ d̪].

2.1.3 Iconicity

Iconicity is the principle according to which a phonetic symbol should resemble in some way the articulation of the sound it denotes, indicating in a direct manner articulatory postures. For this reason, iconic notations are organic (see Section 2.1.1) and also analogical because the same articulatory features will always be represented in the same way, as seen in Section 2.1.2 in relation to Bell's *Visible Speech* notation. Iconicity can vary considerably in how pictorial a symbol is. In addition to his more abstract analogical notation described in Section 2.1.2, Wilkins (1668)

presented another chart consisting of portraits of a speaker's head and neck showing, sometimes in cutaways, the positions of the speech organs for consonants and vowels. Accompanying each one, besides a letter or digraph functioning as a phonetic symbol, is a more diagrammatic representation of lips, tongue, and palate in positions for stops and continuants. It could be said, however, that these play the role of definitional illustrations rather than symbols; it would be very difficult to write transcriptions with them, and this brings up a key difficulty with iconic notation: the more accurate it is pictorially, the more cumbersome it is to use. Conversely, the less accurate it is, the more it has to be explained in conventions, thus undermining the motivation for iconicity. It is not self-evident, for example, that Bell's [**I**] symbol denotes voice, or that his [ˢ] denotes nasality. While it can be useful to draw attention to iconic symbols when teaching phonetics, they are not serious contenders for use in phonetic transcriptions.

Although fully iconic notation is no longer seen as something to aim for, iconicity has nonetheless a presence in the current IPA notation among the diacritics and tone letters. Clear examples are the voiceless diacritic [◌̥] (reminiscent of Bell's 'glottis open' symbol [O]), the dental diacritic [◌̪] in the shape of a tooth, and the linguolabial diacritic [◌̼] iconic of an upper lip.

Because sound frequencies, including fundamental frequency, can be described as inhabiting an auditory high-low dimension, the pitch of the voice can readily be represented by its relative position in a visual analogue. The IPA level-tone letters [˥ ˦ ˧ ˨ ˩] (= extra high, high, mid, low, extra low) and contour-tone letters [˩˥ ˥˩ ˦˥ ˩˨ ˦˥˦ ˧˩˧] (= rising, falling, high rising, low rising, rising-falling, falling-rising), designed and introduced by Chao (1930), exploit this audio-visual parallel iconically.

2.1.4 Componentiality

Phonetic analysis identifies the component features of speech sounds. A type of notation that presents the componential content of sounds without the use of symbols is *analphabetic* notation, pioneered by Otto Jespersen (Jespersen, 1889). Greek and Roman letters for active and passive articulators, and numerals for degrees of stricture, are combined to express the phonetic properties of a consonant or vowel. A tense labiodental fricative (IPA [f]) is represented as *αe1*, as explained in (1).

> (1)
>
> *α* (= lower lip) *e* (= dental) *1* (= close approximation) (italic face = tense)

Pike (1943) pushed the analphabetic componentiality principle to an extreme such that upward of 30 upper-case and lowercase Roman letters in Roman and italic faces were strung together to express all the phonetic features and properties that Pike identified as components of a consonant or vowel; his formula for IPA [t] is given in (2).

> (2)
>
> *M*aIlDe*C*VveIcAP*pa*a*td*tl*n*ra*n*sfsSi*F*Ss (Pike, 1943, p. 155)

Pike's system is hierarchical. The interpretation of a letter in the string depends on which italic letter dominates it from the left. For example, 'a' means 'airstream mechanism' when dominated by '*M*' (= productive mechanism), but 'alveolar' when dominated by '*p*' (= point of articulation). Analphabetic notation was never meant for running transcriptions; its value is in its ability to express a highly detailed analysis of a specific sound-type.

2.1.5 Integralness

A symbol is *integral* if it has a one-to-many relationship with the categories it denotes. For example, the IPA symbol [z] denotes the categories 'voiced, alveolar, fricative' but, in contrast to analogical symbols, no part of the symbol can be identified as denoting any one of the categories separately. Recruiting alphabetic letters as phonetic symbols results in integral symbols because the original letters were not designed to denote discrete phonetic categories. Insofar as letters in writing systems are phonographic, they correspond to sounds in an unanalyzed holistic manner, although the Korean Hangŭl (also spelled 'Hankul') writing system is an exception (King, 1996) and has been adapted for phonetic transcription (Lee, 1999). Because most of the symbols in the basic IPA stock derive from alphabetic letters, they are integral by virtue of their historical origins.

2.2 IPA Notation

In the 1870s, Henry Sweet developed what he called 'romic' notation based on Ellis's palaeotype but with fewer digraphs and no uppercase letters. Sweet aimed to exploit the familiarity of Roman letters by using them for their original sound values, but they had to be augmented by more diacritics, more turned versions of Roman letters, and letters from Greek and other alphabets. Sweet's romic came at a time when the confluence of interest in spelling reform, foreign language teaching, and a desire for an international language that had marked the English School in the 16th and 17th centuries, was again evident. Indeed, Sweet himself was a spelling reformer. English-language teachers in Paris led by Paul Passy founded *L'Association Phonétique des Professeurs Anglais* in 1886 (becoming in 1897 the *International Phonetic Association*) with the express purpose of introducing phonetic notation into teaching materials, and the Esperanto movement's first publication appeared the following year (Cresswell & Hartley, 1957, p. 9), expounding the virtues of phonetically motivated spelling and the 'one letter, one sound' principle. These conditions were conducive to establishing a phonetic notation for use by language teachers based on Sweet's romic, thus the IPA came into being, the first IPA chart appearing in 1889. From its beginnings, its aim was to provide symbols for "each distinctive sound; that is, for each sound which, being used instead of another, can change the meaning of a

word" (IPA, 1999, p. 196). This focus on phonemic function has weakened somewhat as phonetics has become less pedagogical and more scientific, with a concern to be able to express as much as possible of what can be observed, including instrumentally. This concern can be seen at its fullest in the ExtIPA and VoQs systems described in Sections 2.3.1 and 2.3.2.

IPA notation is basically integral but, as pointed out in Section 2.1, makes use of analogy and iconicity as part of its method of expressing the componential structure of speech sounds. This method takes a set of symbols, recruited mainly from the Roman alphabet, known to represent commonly occurring sound-types across the world's languages and modifies them with diacritics to specify properties not expressed by the basic symbol. For example, [t] denotes a 'voiceless alveolar plosive', a consonant type found in many languages, but in some languages, there is an extra modification that can be represented by adding a diacritic to give aspirated [tʰ] as in English *tea*, palatalized [tʲ] as in Russian *teni* 'shadows', and pharyngealized [tˤ] as in Arabic *ṭifl* 'child'.

The set of symbols and diacritics are set out on a chart that itself is to a certain extent iconic. Places of articulation on the main consonant chart are presented left to right from bilabial through to glottal. Manners of articulation are given top to bottom from maximally constricted to maximally open. The vowel chart has front vowels to the left, close vowels at the top. The conventions for interpreting the symbols are supplied by the category labels and by left-right pairings of symbols for voiceless and voiced consonants, and unrounded and rounded vowels. What the category labels mean is the province of phonetic theory, giving some scope for alternative interpretations. For example, the term 'voiceless' can be understood as lacking vocal fold vibration, or as produced with vocal fold abduction, a difference with implications for sounds with glottal constriction but no voicing such as a glottal stop. There have been different views on the status of the chart's category labels, with Abercrombie (1967, p. 124) saying they provide "rough general phonetic definitions for the symbols," but the IPA (1999, p. 159) saying in its Second Principle that a symbol "is a shorthand way of designating the intersection of the categories." The IPA view is the more scientific in that the categories of phonetic theory exhaustively define the symbols instead of merely roughly defining them. It should be remembered, though, that exhaustive categorization is not the same as exhaustive description.

The IPA chart has been periodically revised and expanded in response to a number of factors. Symbols for sounds newly encountered need to be incorporated, for example, the introduction of [ⱱ] for a labiodental flap in 2005, the design being motivated by [v] and [ɾ]. Symbols may be changed because of requests, as happened in 1989 when the click symbols [ʗ ʘ ʖ] were replaced by [ǀ ǃ ǁ]. The location of classes of sounds may be changed in line with developments in phonetic theory, as happened in the early years of the IPA with sounds currently classed as pharyngeal. They first made an appearance in the 1899 chart as 'guttural' and were 'bronchial' in the 1905 version before a 'pharyngeal' column was added in 1926 *<https://www.internationalphoneticassociation.org/IPAcharts/IPA_hist/IPA_hist_2018.html#10>*. Even today there is some unease at their pharyngeal classification in light of instrumental evidence of epiglottal and aryepiglottic involvement (Esling, 2010, pp. 695–697).

An interactive updated IPA chart can be found on the International Phonetic Association website *<https://www.internationalphoneticassociation.org/IPAcharts/inter_chart_2018/IPA_2018.html>*. Clicking on a symbol brings up recordings of that sound by four well-known phoneticians. A Braille version of the IPA is presented in Englebretson (2009), which has a print version of the IPA Braille chart.

Linguists researching the Indigenous languages of the Americas developed their own alternatives to IPA symbols known as the American Phonetic Alphabet (APA), used, for example, in Odden (2005). Commonly encountered APA symbols are [š ž č ǰ], equivalent to IPA [ʃ ʒ ʧ ʤ].

2.2.1 ExtIPA Notation

The set of symbols, diacritics, and conventions known as the Extensions to the IPA (ExtIPA) has been assembled for dealing with sounds that occur specifically in disordered speech. It evolved from the *Phonetic Representation of Disordered Speech* symbols through the work of a group set up with IPA approval and was first presented in Duckworth et al. (1990). The ExtIPA chart was revised extensively in 2015 (Ball, Howard, & Miller, 2018) and can be found on the International Phonetic Association website *<https://www.internationalphoneticassociation.org/sites/default/files/extIPA_2016.pdf>*.

The layout of the ExtIPA chart is modeled on the IPA chart with the additional place categories labio-alveolar, dento-labial, bidental, inter-dental, and velo-pharyngeal, and the new manner categories lateral+median fricative (for simultaneous lateral and median airflow), nasal fricative, and percussive. The place category linguo-labial, denoted on the IPA chart by a diacritic, is promoted to the main chart. A striking difference between the IPA and ExtIPA charts is the extensive presence on the latter of diacritics denoting place and manner categories and voicelessness. There are new diacritics, and new "other sounds" that cannot readily be fitted onto the main chart. A section that has no counterpart on the IPA chart is headed "Connected speech, uncertainty, etc.," showing ways of representing pause length, speech tempo, loudness, and degrees of indeterminacy. A section is devoted to voicing that extends the IPA voiceless and voiced diacritics [◌̬ ◌̥] to show degrees of pre- and post-voicing and devoicing, which can probably be reliably identified only instrumentally.

New symbols are derived by turning and reversing existing symbols, for example, 'voiced pharyngeal plosive' (long believed by the IPA to be an impossible articulation) is denoted by 'ɢ' turned through 180°, and velodorsal oral and nasal stops are denoted by reversed versions of 'k ɡ ŋ'. It is notable that ExtIPA, like the IPA, does not favor the introduction of new glyphs.

2.2.2 VoQS Notation

A notation system for voice quality, known as VoQS, has developed more or less in parallel with ExtIPA. Introduced in 1995, it was revised in 2016 (Ball, Esling, & Dickson, 2018) and can be accessed at the International Clinical Phonetics and Linguistics Association website *<https://www.icpla.info/journal-publications>*. Based on the voice quality descriptions in Laver (1980), and drawing on the conventions of the IPA and ExtIPA, it provides symbols and diacritics for denoting

airstream types, phonation types, and supralaryngeal settings. Using a capital V for ‘voice’, diacritics are added, for example, [V̰] for creaky voice, and [Vˠ] for ‘velarized voice’. Some IPA glyphs are used with a different convention, for example, [!] denotes harsh voice quality. There are also some new diacritics such as turned superscript ‘w’ in [Vᵐ] for ‘aryepiglottic phonation’. The chart shows an example of how VoQS symbols can be used over stretches of speech using curly brackets—see example (7) in Section 3.2.

2.2.3 Specialist Notations

Two specialist notation systems deserve mention. One is a system for representing infant vocalizations, the other for conversation analysis.

A system of notation such as the IPA is not well suited for representing the prelinguistic vocalizations of infants because, due to immaturity, their vocal tracts cannot produce the kind of phonetic output that matches some of the categories underlying IPA notation. However, their vocalizations can be analyzed into recurring acoustic-auditory categories, which can be denoted. Oller (2000, pp. 193–194) calls such sounds ‘proto-phones’ and represents them using abbreviations of appropriately descriptive category terms, for example, [QNR] = ‘quasi-resonant nucleus’, [SQ] = ‘squeal’, [GR] = ‘growl’, and [GO] = ‘goo’.

The notation that has been developed by practitioners of conversation analysis (CA) aims to represent phonetic details thought to be important for understanding the structure of conversational interaction (Jefferson, 2004). It recruits orthographic resources, giving them specific conventions. Phonographic respellings are used rather than proper phonetic notation. Walker (2013) offers a critique of CA notational practices and shows how CA transcriptions can be enriched by complementary use of other phonetic notations, including the IPA.

2.3 Prosodic Notation

In this article *prosody* refers to those features of speech that are not regarded as inherent to consonants or vowels. These include such features as rhythm, intonation, tone, accent, and stress. Phonographic writing systems have never developed the means of representing these aspects of speech beyond a few diacritics for accentual pitch, for example, the Greek grave, acute, and circumflex signs, and the tone signs of Lao and Thai writing. Prosodic notation. therefore. has had to be developed almost entirely independently of alphabetic influences, so it is not surprising that linguists have employed different systems. According to Wells (2006, p. 261), some systems “are mere notational variants . . . others are based on different theoretical assumptions.” While there is little room for disagreement over what constitutes a consonant or vowel, there is scope for differences of opinion over what kinds of prosodic units need to be recognized in speech and how they should be symbolized.

Notations for lexical tone, stress, rhythm, and intonation will be briefly discussed while remaining as neutral as possible in relation to theoretical assumptions and terminological usage.

2.3.1 Notation for Lexical Tone

Lexical tone is the use of pitch to distinguish between words of the same segmental structure. Languages that do this are known as *tone languages* and represent more than half of the languages of the world. Lexical tone can be notated in several ways. The iconic IPA tone letters have already been presented in Section 2.1.3, but the IPA provides diacritical alternatives that are less iconic but easier to use in transcriptions. For convenience, some linguists prefer the integer equivalents of the Chao tone letters, in which level tones are numbered 1–5 from lowest to highest (which Chao suggests encompass five whole musical tones), and contour tones are denoted by combining two level-tone integers, for example, 13 for a rise from lowest to mid. Integers are also used to arbitrarily identify a particular tone in a specific language. In Mandarin Chinese, for example, tones are numbered 1–4, in Thai 1–5. Similarly, in Swedish the two contrasting pitch accents are numbered 1 and 2. The arbitrary numbering of tones and pitch accents is an otherwise rare example of phonetic classification terms that have no descriptive content. Examples of alternative notations for a falling tone are given in (3) with the syllable [ma] ([k^ha:] for Thai).

> (3)
>
> IPA tone letter: ma˥˩
>
> IPA tone diacritic: mâ
>
> Pitch level integers: ma^{51}
>
> Pinyin (for Mandarin): mà ‘to scold’
>
> Tone number (for Mandarin): ma^4 ‘to scold’
>
> (for Thai): k^ha:^{3} ‘value’

While diacritics imply that tone is a property of vowels, the placing of tone letters and integers after the syllable suggests it is a property of whole syllables. Phonologists take different views on the relevant domain of lexical tone and may choose notation accordingly.

For languages with only two or three level tones, they can be denoted as ‘H’ for high, ‘M’ for mid, and ‘L’ for low, with the letter placed above the syllable as in the example from Nupe in (4).

(4)

H	M	L
[ba] ‘to be sour’	[ba] ‘to cut’	[ba] ‘to count’

In some tone languages, a tone has an extra contextually determined up-step or down-step in pitch which can be represented by placing a superscript up or down arrow before the syllable, for example, [má$^{\uparrow}$má] for an up-step, [mà$^{\downarrow}$mà] for a down-step.

2.3.2 Notation for Rhythm

For primary and secondary word-stress, the IPA provides superior and inferior vertical strokes, respectively. They can be used with orthographic forms as well as in phonetic transcriptions, for example, ˌ*mathema*ˈ*tician*, and are now the favored stress marks in most dictionaries. When dealing with rhythmic stress in stretches of continuous speech beyond single words, rhythm group (foot) boundaries can be shown by a thick vertical line as in (5).

(5)

they’ve | taken it to the De | partment of E | lectrical Engi | neering |

A notational innovation for representing an analysis of rhythm is the metrical grid (Prince, 1983). Asterisks are placed over syllables, the number of asterisks indicating the relative rhythmic prominence of the syllable, as shown in (6).

(6)

```
                    *
      *             *
      *     *       *
 *  *  *  *  *  * * * *
an eccentric mathematician
```

An advantage of a grid notation is that it is easy to see which syllables in a long string have the same degree of prominence and, when they occur adjacently on the same level, whether there is an intervening syllable with non-minimal prominence to prevent ‘stress-clash’.

2.3.3 Notation for Intonation

The IPA provides double thick vertical lines for marking intonation group boundaries ‖. . .‖, but it provides precious little for marking anything else connected with intonation beyond the rather vague 'global rise' [↗] and 'global fall' [↘], as the IPA itself acknowledges (IPA, 1999, p. 14). Writers on intonation have therefore often presented their own particular notations and accompanying conventions. Many have in common, however, that they use marks iconic of pitch movement during the nuclear tone. Falling tones are represented by downward-sloping marks, rising tones by upward-sloping marks, and falling-rising and rising-falling tones by marks that change direction accordingly: [ˋ ˊ ˇ ˆ], the marks being placed before the tonic syllable; low falls and rises can be shown by lowering the position of the tone mark: [ˎˏ].

Starting with Jones (1972, pp. 277–324), and associated more with the British phonetics tradition, is *interlinear tonetic transcription*, which uses a notation consisting of a two-line stave representing the speaker's normal pitch range, with dots of different sizes to indicate the relative degree of prominence of each syllable, some with tails attached to indicate the direction and extent of the tone's pitch movement. The dots are placed on the stave to indicate the pitch contour of the utterance.

Instead of explicitly representing the dynamics of pitch movement with iconic notations, many intonational phonologists prefer a notation that denotes fixed relative pitch levels. A system that has gained currency over recent decades is the ToBI system (Tones and Break Indices; Beckman et al., 2005). Its symbols and conventions are given in (7).

> (7)
>
> H = high; M = mid; L = low; * = accent; ! = stepped accent (e.g. in a sequence H* !H* the second accent is stepped down from the first); - = phrasal accent/tone; ^ = extra high; > = displaced; % = boundary; bitonal pitch accents are conjoined with +, for example, L+H* = a high accent with a low leading.tone

Figure 2 compares an interlinear tonetic with a ToBI transcription.

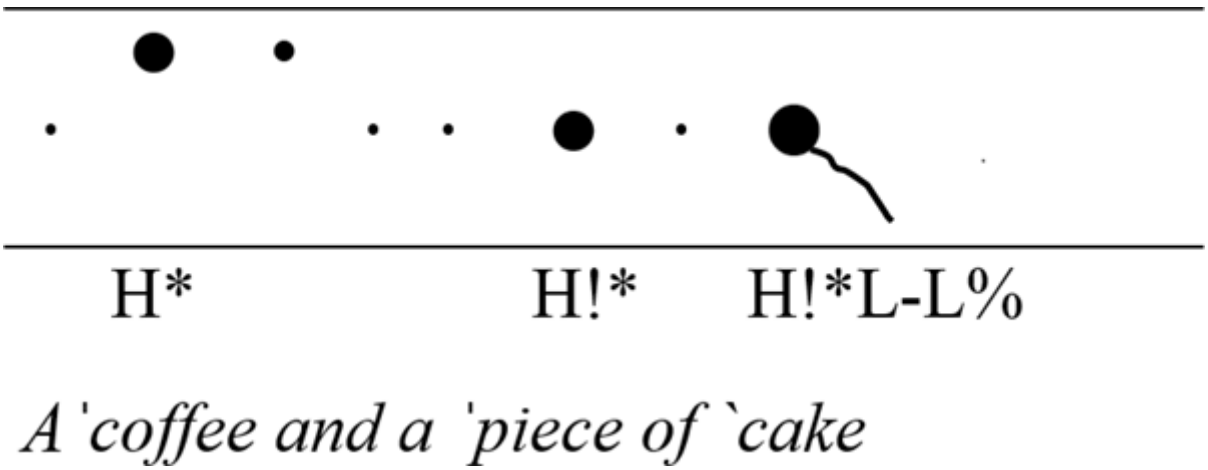

Figure 2. ToBI and interlinear tonetic transcriptions of *A coffee and a piece of cake,* and orthographic transcription with stress and tone marks (see also figure 4).

Hualde and Prieto (2016) discussed shortcomings in the extent to which ToBI notation can mean the same in transcriptions of different languages. They proposed trying to make the system more universal by having broad phonetic intonational categories in addition to more abstract phonological ones, thus in effect making it a notational variant to the interlinear tonetic system. A system using a similar notation to ToBI but designed explicitly to represent universal surface phonetic intonational phenomena is the INTSINT system (Hirst, 2011). For any utterance, the fundamental frequency contours are calculated from the acoustic signal, and a set of eight acrophonic symbols are then used to specify the intonational melody at crucial points. Three have absolute Hz values: T (= top of the pitch range), M (= mid-point of the range), B (= bottom of the range); two have relative values: H (= high) and L (= low); and two have iterative values: U (= upstep) and D (= downstep) (Hirst, 2011, p. 70).

3. Transcription

Phonetic transcription is the deployment of phonetic notation to represent an analysis of the phonetic content of actual or potential utterances. Common places to encounter phonetic transcriptions are dictionaries, language teaching and learning materials, linguistic descriptions of languages and language varieties, materials for learning and teaching phonetics, and phonetic analyses of speakers' behaviors, whether for speech research, or for clinical or forensic purposes. As explained in Sections 3.1 through 3.6, there are different types of transcriptions depending on what the transcriber's purposes are. Transcriptions can also have different functions. A transcription can be a (purportedly) objective descriptive record of an utterance, or it can have a prescriptive function in providing a model for pronunciation as, for example, in a dictionary for language-learners. Transcriptions can also be performance scores in dictation and production exercises for students. An increasingly common use for transcription is for the aligned annotation of instrumental records such as sound spectrograms and waveforms, and palatograms; this is discussed further and exemplified in Section 3.6.

3.1 Segmental and Parametric Transcription

It is taken as axiomatic in phonetics that running speech is not produced in a 'one-sound-at-a-time' fashion, and that division of speech into discrete consonant and vowel segments is a somewhat artificial procedure (Laver, 1994, pp. 566–568). Segmentation is, however, a highly useful procedure for phonetic and phonological analysis, and the notation most favored for making segmental transcriptions is the IPA. That speech production involves speech organs continuously moving in and out of play in a complex pattern of coordination can be shown schematically in *parametric transcriptions* (Tench, 1978). An example is shown in Figure 3. Lateral movements, however, are not easy to represent, and the difficulty in making and reading parametric transcriptions renders them unsuitable for most purposes; however, constructing them is a useful student learning exercise. A more abstract type of parametric representation is a *gestural score*, which is central to theories of phonology that take articulatory gestures as their primes (e.g., Gafos, 2002).

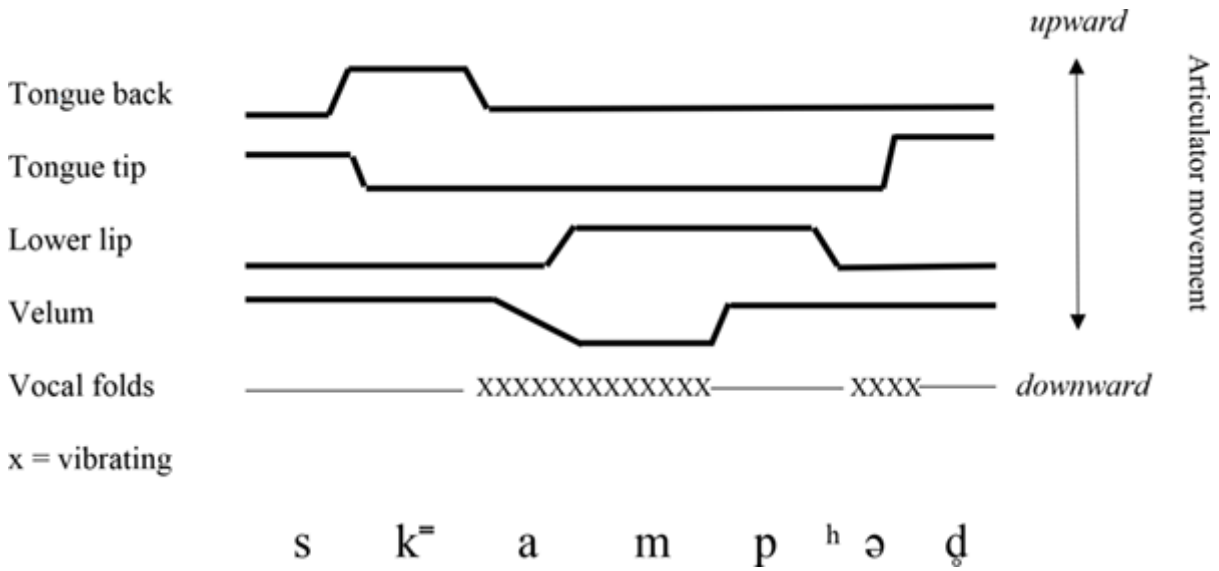

Figure 3. Parametric transcription of *scampered* with aligned segmental transcription.

Source: Adapted from Heselwood (2019, p. 1360).

3.2 Broad and Narrow Transcription

The terms *broad* and *narrow* for transcription were first coined in Sweet (1877), although essentially the same distinction is to be found in Ellis's terms *approximative* and *complete* (Ellis, 1867). To prevent confusion with the *systematic* versus *impressionistic* and the *phonemic* versus *phonetic* distinctions, the broad versus narrow distinction can be based squarely on the amount of phonetic detail contained in a transcription. Broad transcriptions will tend to have consonant and vowel symbols with few diacritics, while narrow transcriptions are likely to have several diacritics to indicate more precisely places and manners of articulation, vowel qualities, and voicing, and

perhaps also voice qualities and dynamics. Three transcriptions of the same utterance of the English word *keys* are given in (8), (8a) being the broadest, (8c) the narrowest in the sense of denoting the most categories.

> (8)
>
> a) [kiːz]; b) [kʰiːz̥]; c) [{Vʲk̟ˣʰiːz̥ˠVʲ}]

Narrowness is not, however, simply a matter of counting diacritics. The Roman alphabetic basis of the IPA, and its historical leanings toward English, French, and German, are largely responsible for whether certain categories are denoted integrally or by diacritics. Voiceless nasals, for example, are denoted by the addition of a diacritic but voiced nasals are not, although the same number of phonetic categories are present in both [m̥] and [m]. It is therefore the number of phonetic categories denoted in a transcription, not the number of diacritics, that should define how broad or narrow it is.

3.3 Systematic and Impressionistic Transcription

An impressionistic transcription is made from the sense-impressions of the transcriber, which may be audiovisual or just audio. The transcriber attends to the utterance, tries to recognize the consonants and vowels, accents, and pitch movements, and records them in the transcription. In an impressionistic transcription, little or no recourse is made to knowledge of the phonological system of the language being spoken. Linguists undertaking fieldwork with speakers of languages that have not yet been analyzed will necessarily make impressionistic transcriptions, as will speech and language clinicians working with clients whose difficulties with speech affect their pronunciation in unpredictable ways. Dialectologists, sociophoneticians, and forensic phoneticians will often also want to make transcriptions uninfluenced by their prior knowledge of the language. The extent to which it is possible to keep such influences out is, however, debatable (Ladefoged, 1990, pp. 340–341).

In the case of systematic transcriptions, details can be left out if they are supplied by language-specific conventions. For this reason, systematic transcriptions are typically broader than impressionistic ones. A systematic transcription of *keys* need only contain the information in (8a). It is known from previous studies of English that a voiceless plosive will be aspirated in this context, and that voiced obstruents are typically devoiced when not followed by another voiced sound.

There has to be scope in systematic transcriptions for including variant realizations that cannot be predicted from the conventions; they are, therefore, not necessarily phonemic. For example, which allophone of /t/ will occur in a word such as English *letter* is not predictable from knowledge of the phonological system: speakers can use any of (at least) [t ʔ ɾ]. A systematic transcription [ˈlɛʔə] is not claiming that [ʔ] is a phoneme.

3.4 Transcription Orientation

Instrumental evidence reveals that what a listener hears can be at variance with what the speaker is doing. This mismatch has obvious implications for impressionistic transcription. Research reported in Hillenbrand and Houde (1996) showed that listeners are likely to hear a glottal stop even if the speaker is continuing to phonate provided that there is a sufficiently sudden drop in amplitude and fundamental frequency. The question then arises of whether a glottal stop should be transcribed. A speaker-oriented transcription would not include a glottal stop symbol, but a listener-oriented transcription would. Howard and Heselwood (2011) presented and discussed further examples from typical and atypical speech of articulatory activities revealed instrumentally that suggest something different from what transcribers hear. A particularly instructive example is of an adult speaker with severe apraxia saying *a jaw*, which was transcribed impressionistically as [ə bːɔə]. Electropalatographic evidence showed that a postalveolar closure for [ʤ] was made while the bilabial closure was in place, but no [ʤ] was heard (Howard & Heselwood, 2011, pp. 944–945). As a way of dealing with such cases, they proposed *two-tier transcriptions* that express both the perceptual analysis and the instrumental analysis, and argued that neither should be seen as intrinsically more correct than the other; they should instead be seen as complementary perspectives.

3.5 Phonemic and Allophonic Transcription

Once a phoneme inventory has been established for a language variety, a symbol can be assigned to each phoneme. A transcription that uses only those symbols is called a phonemic transcription and is enclosed in forward slashes /. . ./. Rather than representing a surface analysis of the pronunciation of words, a phonemic transcription is a statement of their phonological form. Taking a phoneme to be a set of sounds that have the same distinctive function in the language in question, it follows that a symbol in a phonemic transcription denotes that whole set even though, in any particular context, only a subset could actually occur. Which member or members of the set can occur in a context is recoverable from the phonological rules of the language that function as conventions for phonemic transcriptions.

The sounds comprising a phoneme are called the *allophones* of that phoneme. A transcription that represents those allophones is an allophonic transcription. The difference between a phonemic and an allophonic transcription is similar to the difference between a broad and a narrow phonetic transcription in that more phonetic detail is contained in an allophonic transcription. There is, however, a crucial difference, which is that while a symbol in a broad phonetic transcription denotes a single sound-type, in a phonemic transcription it denotes a set of sounds. For example, in a broad transcription of English *letter* as [ˈlɛtə], [t] represents a voiceless alveolar plosive, whereas in the phonemic transcription /ˈlɛtə/ it denotes a set that includes sounds that are not voiceless ([ɾ]) and not alveolar ([ʔ]). It is a common mistake among those with an incomplete grasp of the concept of the phoneme to think of a phoneme symbol as representing a single sound-type.

Symbols are assigned to phonemes typically by identifying a phoneme's principal allophone and removing any unnecessary diacritics. The principal allophone is generally taken to be that which occurs as a singleton in the onset of a stressed syllable. For English voiceless stops, these are [p^h t^h k^h], but because aspiration is predictable by phonological rule, the unadorned symbols /p t k/ are used. In Korean, however, aspiration is not predictable and is the basis of the phonological contrast in word pairs such as [pal] 'sucking' and [p^hal] 'arm' (Lee, 1999, p. 120). Korean therefore requires the phoneme-symbol /p^h/ as well as /p/.

3.6 Transcriptions Aligned With Instrumental Records

The results of instrumental analyses of speech can be displayed visually. To assist in their interpretation, transcriptions can be temporally aligned so that it can easily be seen what the display is revealing about each consonant and vowel, or each rhythmic or intonational event. Figures 4 and 5 provide examples.

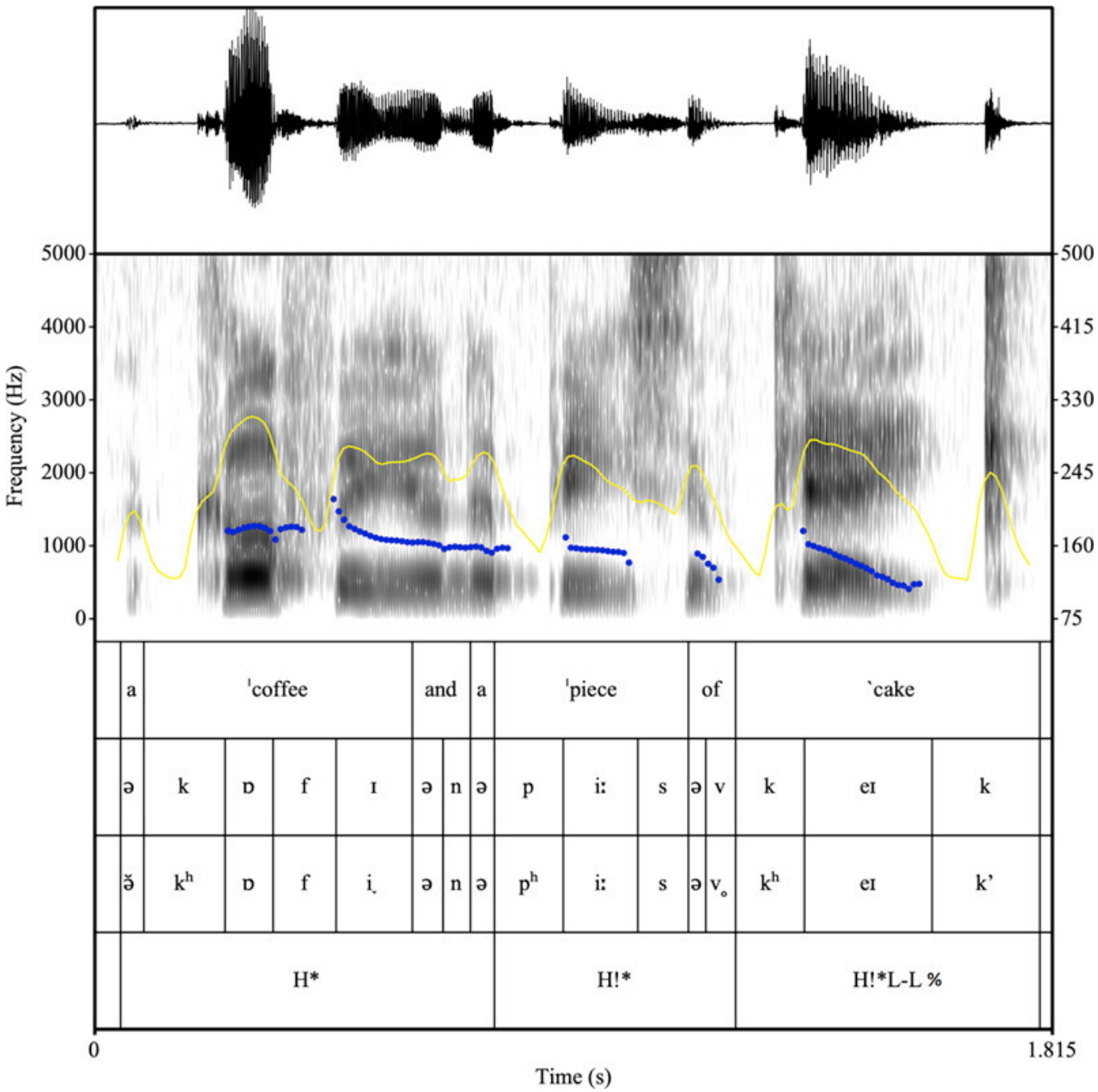

Figure 4. Spectrogram and waveform of *A coffee and a piece of cake* with pitch (lower) and intensity (upper) traces. Tier 1: orthographic transcription with stress and tone marks; Tier 2: phonemic transcription; Tier 3: broad allophonic transcription; Tier 4: ToBI intonational transcription.

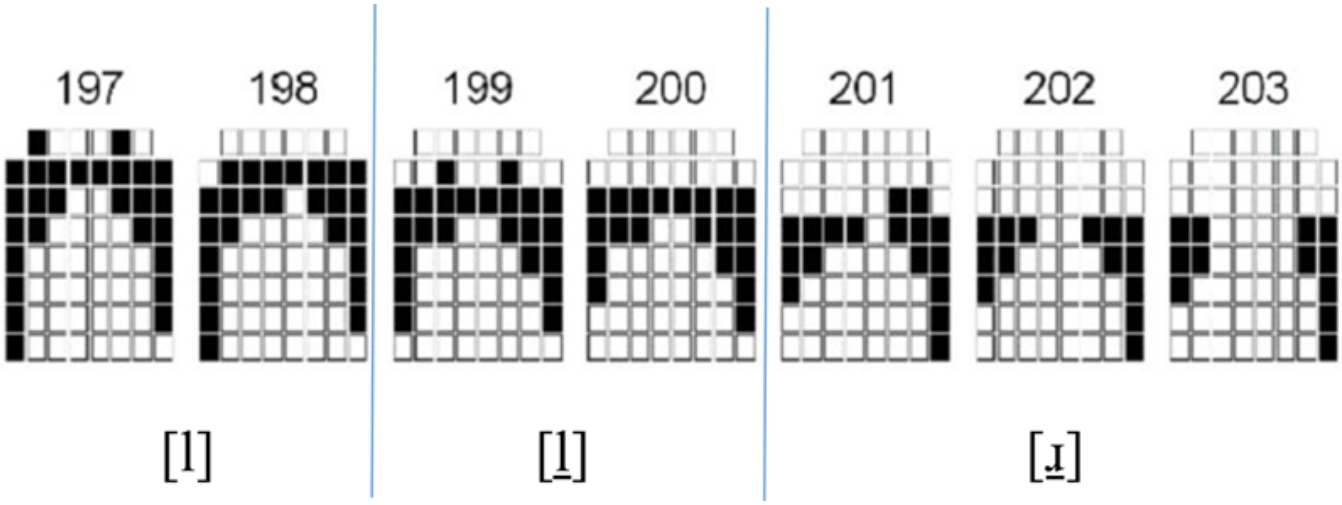

Figure 5. Palatogram frames showing articulatory transition from [l̪] through [l̪] to [ɹ̪] in the Arabic phrase /ˈħabil raˈfiːʕ/ ‘a thin rope’.

Transcriptions aligned with instrumental records can perform two different functions. Firstly, they help to identify where the various pieces of linguistic material are located. In Figure 4, the symbols indicate where particular vowels and consonants are in the spectrogram and waveform. Furthermore, the segment boundaries enable automatic measurement of variables such as segment durations, vowel formants, and pitch through the application of scripting—tutorials on scripting for the Praat phonetic analysis software can be accessed at the Praat Scripting Tutorial website *<http://praatscripting.lingphon.net/index.html>*, the software itself is free from “Praat: Doing Phonetics by Computer *<http://www.fon.hum.uva.nl/praat/>*”. Secondly, transcriptions can help to interpret what the instrumental analysis is saying about that material. For example, the [k’] symbol in Figure 4 tells the reader that the spectrographic display provides evidence for an ejective realization of /k/; in Figure 5 the [l̪] draws attention to the palatographic evidence for dynamic retraction of the articulatory contact during the realization of /l/ under the influence of a following /r/.

4. Issues in Transcription

A recurring issue in phonetic transcription concerns the value of impressionistic transcription. Because impressionistic transcriptions of the same data can vary, the often-leveled charge of unreliability is irrefutable if reliability means getting the same results from the same data using the same methods. However, impressionistic transcriptions can be useful and insightful despite not meeting strict criteria of reliability, the more so if certain procedures are followed. Chief among these is working from good-quality recordings. Trying to transcribe “live” while the speaker is talking is likely to lead to much lower reliability because transcription cannot keep up, and the consequent tendency to normalize the transcription (Amorosa et al., 1985) will lead to less insightfulness. In order to avoid top-down influences, transcribers should not know in advance the target utterance. Reliability has been found to increase if transcriptions are compared by aligning them intelligently and using distance matrices that try to quantify similarities

between sounds (Cucchiarini, 1996), not just by counting symbol agreement. To tackle the issue of reliability, Shriberg et al. (1984) suggested procedures for arriving at *consensus transcriptions* from transcriptions of the same data made independently by different transcribers. The validity of consensus transcriptions, however, has been questioned insofar as they express analyses that none of the transcribers made (Heselwood, 2013, pp. 218–219).

As instrumental research reveals more about phonetic structure, including phenomena beyond human perceptual abilities, it may become necessary to find notational means for denoting them. Or a limit may be reached as to what can usefully be represented in transcriptions. For example, should the different VOT range-values that characterize stops in different languages (Cho & Ladefoged, 1999) be symbolized, especially those that may fall below difference limens? As so often in the history of transcription, it will likely be its practitioners who decide.

A further issue relates to the ontology of the denotata of phonetic symbols insofar as they are said to be "speech sounds." What kind of an object can a speech sound be said to be—a complex of articulatory postures, a bundle of muscular and aerodynamic events in the vocal tract, a pattern of acoustic disturbances, or an auditory quality? The terms used in phonetic notation conventions owe a heavy debt to the anatomy of the vocal tract, but the choice of which phonetic symbols to use in a transcription is commonly made as a result of auditory judgement. Whatever kinds of objects speech sounds may be, it is not altogether clear that they can be delimited in the stream of speech in an objective and principled manner. This issue perhaps does not need to be resolved in order for successful phonetic transcription to take place, yet there is something unsatisfactory about leaving the notion of a speech sound undefined when it is such a central notion in the theory and practice of phonetic transcription.

5. Transcription in Electronic Speech Corpora

An electronic or machine-readable speech corpus is a database of transcribed, spontaneous and/or read speech, accessed via speech audio files (.wav format) and other file formats, that can be processed by a computer without human intervention. One such database is the acoustic-phonetic *TIMIT* corpus of continuous speech (Garofolo et al., 1993), which complements each phonetically rich, recorded utterance with its orthographic transcription, plus word and phonetic transcriptions as vertical time-stamped events.

Annotated speech corpora are essential resources for two major Natural Language Processing tasks: Automatic Speech Recognition (ASR) and Text-to-Speech (TTS). The former entails automatic and accurate transcription (or 'recognition') of human speech into/as text, while the latter automates the transformation (or 'synthesis') of text into naturalistic-sounding speech. Both technologies address the same goal in artificial intelligence: to enable and enhance human-computer interaction through the medium of natural language. Furthermore, both technologies depend on phonemic and/or phonetic transcription to mediate between the spoken and written word. Thus, in ASR a transcription such as /bʊk/ acts as an intermediary symbolic and linguistic representation between the speech signal and the output <book>, while in TTS it represents a symbolic grapheme-to-phoneme encoding of the word token <book> prior to speech synthesis.

Speech corpora used in ASR and TTS for English feature in the Linguistic Data Consortium's list of the top 10 most widely distributed corpora and were created in the 1990s; these include the previously mentioned TIMIT, and *Switchboard-1 Release 2* (Godfrey & Holliman, 1993). In particular, the Switchboard data set of conversational telephone speech has consistently been used to benchmark performance of ASR systems, including deep learning and end-to-end neural network approaches for state-of-the-art speech recognition (Xiong et al., 2017). However, a recent and alternative approach to using well-known corpora is mobile phone capture of large amounts of data from many speakers in a variety of recording environments. For example, Hughes et al. (2010) compiled a corpus of high-quality read speech where prompts and text transcripts are sourced from web queries featuring nonstandard vocabulary items.

5.1 Alignment in Electronic Speech Corpora

The procedure for phonetic transcription and alignment in TIMIT is described in Zue and Seneff (1996). The speech waveform was aligned automatically with the manually entered, acoustic-phonetic input string via bespoke software, developed at the Massachusetts Institute of Technology, and reported to have correctly performed over 95% of the labeling task previously carried out by human transcribers. Automatically generated boundaries were then scrutinized and corrected by an acoustic phonetician.

TIMIT is a corpus of read speech containing 630 speakers representing 8 major dialects of American English each reading 10 sentences. In the Switchboard corpus of conversational telephone speech, 72 minutes (out of 260 hours) of recorded speech were phonetically transcribed, using a variant of TIMIT's ASCII-based ARPAbet symbol set: for example, the ASCII character sequence /A:ft@/ equivalent to /ɑːftə/ in IPA notation. Again, the transcription procedure involved human scrutiny and correction of automatically aligned segments and segmental boundaries generated via speech analysis and processing tools.

More recently, automatic, open-source alignment tools based on the *Praat* speech analysis software package (Boersma & Weenink, 2019) have been developed. A shared concept is the parallel representation of alignment data in a TextGrid file, which enables manual oversight and correction. The *EasyAlign* Praat plug-in (Goldman, 2011), originally designed for French, English, Spanish, and Taiwanese, generates multiple annotation tiers (at phoneme, syllable, word, and utterance levels) from the sound recording and its corresponding orthographic transcription, formatted as one sentence per line. Automated segmentation is executed via scripts and entails: macro-segmentation at utterance level; grapheme-to-phoneme conversion; and phone segmentation. The aptly named *AlignTool* (Schillingmann et al., 2018), initially developed for German, Dutch, and British English, uses Praat to pre-segment speech onsets and offsets and then carries out forced alignment of transcripts with the speech signal via speech recognition. The availability of a text transcript, plus grapheme-to-phoneme conversion to create a phonotypical representation of the input string (i.e., the orthographic form), facilitates statistical acoustic modeling of the sounds that make up each word.

5.2 Approaches to Automatic Phonetic Transcription

Grapheme-to-phoneme conversion in AlignTool draws on a phonetic lexicon to generate phonotypical representations of orthographic sequences aligned with the speech signal in previously detected speech intervals. Lexicon look-up is one of three generic approaches to automated phonetic transcription, the others being *data-driven* and *knowledge-based*. The former (data-driven) implements statistical speech recognition to determine the most probable phonetic label sequence for the observed acoustic data; for example, Liang et al. (2008) incorporated a data-driven approach when transcribing the Buddhist Sutra. The latter (knowledge-based) depends on rewrite rules. Brierley et al. (2016) reported on a verified mapping scheme (rule set) for IPA transcription of the entire text of the Arabic Koran, based solely on the orthographic form. Their procedure included frequency profiling of unigrams and n-grams (i.e., single characters and multiple character sequences) throughout the text to identify specific sequences as compound transcription events. For example, the forms {ؤ ئ إ أ ء} are all collected as unigrams and map to /ʔ/, while أ is collected as a bigram and maps to /ʔun/. They also report on an extended rule set for segmenting each canonical IPA transcription of an Arabic word into a sequence of syllable tokens via a discrete set of consonant-vowel patterns for Arabic {CV, CVV, CVC, CVVC, CVCC} plus automatic assignment of primary stress verified via inter-annotator agreement (Brierley et al., 2019). A knowledge-based approach was also adopted in Ramsey et al. (2014) to generate context-sensitive Speech Assessment Methods Phonetic Alphabet (SAMPA) phonetic transcriptions for Modern Standard Arabic speech; their rule set featured phoneme-to-phone (in addition to grapheme-to-phoneme) conversion to model intra-word and cross-word variation in continuous speech.

Van Bael et al. (2007) trialed 10 different procedures, and combinations of procedures, for automatic phonetic transcription via a standard continuous speech recognizer on read and spontaneous Dutch speech. Like TIMIT and Switchboard, their data set contained a manually verified, phonetically transcribed subset which was used as a gold standard for evaluating system performance. Their study involved the three generic procedures (i.e., lexicon look-up, data-driven, and knowledge-based), two combinations of the same, and five further tests where decision-tree filtering was applied to fine-tune the output. Lexicon look-up, fine-tuned by decision trees, was found to be the best approach; another main finding was that learning intra-word and cross-word variation from an appropriate sample of gold standard transcriptions (i.e., the data-driven approach) gave better results than knowledge-based predictions.

5.3 Speech Corpora With Prosodic Annotation

The *Aix-MARSEC* project (Auran et al., 2004) is an iconic speech corpus of multiple, aligned, phonetic, and prosodic annotation tiers developed for TTS from the Lancaster/IBM *Spoken English Corpus* (SEC) and its machine-readable counterpart, MARSEC (Roach et al., 1993; Taylor & Knowles, 1988). Aix-MARSEC features automatically generated SAMPA transcriptions, which capture some reduced forms plus the linking 'r', as in '*after it took*' rendered in SAMPA as /A:ft@ **r**It tUk/. It also uses the MOMEL-INTSINT algorithm to encode a surface phonological representation of intonation directly from the speech signal via a series of F0 target points in a

given utterance that are classified as either absolute tones (i.e., mid, top, bottom) or relative tones (e.g., higher, lower, same). Finally, prosodic annotation in the corpus includes rhythmic units defined according to two schemes: (a) the *stress foot* and (b) the combination of *narrow rhythm unit* (NRU) and *anacrusis* or unstressed syllable (ANA). For example, the phrase '*almost impossible*' can be realized as two feet: |almostim |possible|; alternatively, the word boundary can be preserved by representing it as two NRUs separated by an anacrusis: |almost im|possible| or [NRU] [[ANA][NRU]].

AixOx (Herment et al., 2014) is a multilayered, English-French learners' corpus of read speech that also implements macro-prosodic, MOMEL-INTSINT annotation, following automatic segmentation via the Speech Phonetization, Alignment, and Syllabification (SPPAS) algorithm (Bigi & Hirst, 2012). SPPAS generates utterance, word, syllabic, and phonemic units, with resulting alignments in TextGrid format, and was originally developed for English, French, Chinese, and Italian. Multilevel, prosodically annotated corpora are also available for Russian and Greek. CORPRES, or the Corpus of Russian Professionally Read Speech (Skrelin et al., 2010), uses a language-specific version of SAMPA; and an annotation system for Greek speech corpora has been derived from the Tones and Break Indices (ToBI) schema (Arvaniti & Baltazani, 2000). Both corpora depend on a significant amount of manual as well as automatic annotation.

Further Reading

Abercrombie, D. (1964). *English phonetic texts*. Faber & Faber.

Albright, R. W. (1958). *The International Phonetic Alphabet: Its background and development*. Indiana University Press.

Alsharhan, E., & Ramsey, A. (2019). Improved Arabic speech recognition system through the automatic generation of fine-grained phonetic transcriptions. *Information Processing and Management*, *56*, 343–353.

Esling, J. H. (2010). Phonetic notation. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed., pp. 678–702). Blackwell.

Hardie, A. (2014). Modest XML for Corpora: Not a standard, but a suggestion. *ICAME*, *38*, 73–103.

Heselwood, B. (2013). *Phonetic transcription in theory and practice*. Edinburgh University Press.

International Phonetic Association. (1999). *The handbook of the International Phonetic Association*. Cambridge University Press.

Kemp, J. A. (1994). Phonetic transcription: History. In R. E. Asher & J. M. Y. Simpson (Eds.), *The encyclopedia of language and linguistics* (Vol. 6, pp. 3040–3051). Pergamon.

Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.

Le, J. (2019). *Deep learning-based automatic speech recognition <https://heartbeat.fritz.ai/the-3-deep-learning-frameworks-for-end-to-end-speech-recognition-that-power-your-devices-37b891ddc380>*.

MacMahon, M. K. C. (1996). Phonetic notation. In P. T. Daniels & W. J. Bright (Eds.), *The world's writing systems* (pp. 821–846). Oxford University Press.

Müller, N. (Ed.). (2006). *Multilayered transcription*. Plural.

O'Connor, J. D. (1973). *Phonetics*. Penguin.

Sweet, H. (1881). Sound notation. *Transactions of the Philological Society*, *18*, 177–235.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press.

Amorosa, H., von Benda, U., Wagner, E., & Keck, A. (1985). Transcribing detail in the speech of unintelligible children. *British Journal of Disorders of Communication*, *20*, 281–287.

Arvaniti, A., & Baltazani, M. (2000). Greek ToBI: A system for the annotation of Greek speech corpora. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhauer (Eds.), *Proceedings of 2nd Language Resources and Evaluation Conference* (pp. 555–562). European Language Resources Association.

Auran, C., Bouzon, C., & Hirst, D. (2004). The Aix-MARSEC Project: An evolutive database of spoken British English. In B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2004* (pp. 561–564). Nara, Japan: (n.p.).

Ball, M. J., Esling, J. H., & Dickson, B. C. (2018). Revisions to the VoQS system for the transcription of voice quality *<http://doi.org/10.1017/S0025100317000159>*. *Journal of the International Phonetic Association*, *48*, 165–171.

Ball, M. J., Howard, S. J., & Miller, K. (2018). Revisions to the ExtIPA chart *<http://doi.org/10.1017/S0025100317000147>*. *Journal of the International Phonetic Association*, *48*, 156–164.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework *<http://doi.org/10.1093/acprof:oso/9780199249633.003.0002>*. In S.A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 9–54). Oxford University Press.

Bell, A. M. (1867). *Visible speech*. Simpkin, Marshall & Co.

Bigi, B., & Hirst, D. (2012). Speech phonetization alignment and syllabification (SPPAS): A tool for the automatic analysis of speech prosody. In Q. Ma, H. Ding, & D. Hirst (Eds.), *Proceedings of Speech Prosody 2012* (pp. 19–22). Shanghai, China: Tongji University Press.

Boersma, P., & Weenink, D. (2019). *Praat: Doing phonetics by computer. Version 6.1.08 <http://www.praat.org/>*

Brierley, C., Sawalha, M., & El-Farahaty, H. (2019). Translating sacred sounds: Encoding *tajwīd* rules in automatically generated IPA transcriptions of Quranic Arabic. In S. Hanna, H. El-Farahaty, & A. Khalifa (Eds.), *The Routledge handbook of Arabic translation* (e-book pp. 46–64). Routledge.

Brierley, C., Sawalha, M., Heselwood, B., & Atwell, E. (2016). A verified Arabic-IPA mapping for Arabic transcription technology *<https://doi.org.10.1093/jss/fgv035>*. *Journal of Semitic Studies*, *61*(1), 157–186.

Chao, Y. R. (1930). A system of tone letters. *Le Maître Phonétique*, *45*, 24–27.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT. *Journal of Phonetics*, *27*, 207–229.

Cresswell, J., & Hartley, J. (1957). *Esperanto*. English Universities Press.

Crystal, D. (1987). *The Cambridge encyclopedia of language*. Cambridge University Press.

Cucchiarini, C. (1996). Assessing transcription agreement. *Clinical Linguistics & Phonetics, 10,* 131–156.

Duckworth, M., Allen, G., Hardcastle, W. J., & Ball, M. J. (1990). Extensions to the IPA for the transcription of atypical speech. *Clinical Linguistics & Phonetics*, *4*, 273–280.

Ellis, A. J. (1867). On palaeotype: Or, the representation of spoken sounds, for philological purposes, by means of the ancient types. *Transactions of the Philological Society*, *12*, 1–52.

Englebretson, R. (2009). An overview of IPA Braille: An updated tactile representation of the International Phonetic Alphabet *<https://doi.org/10.1017/S0025100308003691>*. *Journal of the International Phonetic Association*, *39*, 67–86.

Esling, J. H. (2010). Phonetic notation. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The handbook of phonetic sciences* (2nd ed., pp. 678–702). Blackwell.

Firth, J. R. (1946). The English school of phonetics. *Transactions of the Philological Society*, *45*, 92–132.

Gafos, A. (2002). A grammar of gestural coordination. *Natural Language & Linguistic Theory*, *20*, 269–337.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus* *<https://catalog.ldc.upenn.edu/LDC93S1>*. Web download. Philadelphia, PA: Linguistic Data Consortium.

Godfrey, J. J., & Holliman, E. (1993). *Switchboard-1 Release 2* *<https://catalog.ldc.upenn.edu/LDC97S62>*. Web download. Philadelphia, PA: Linguistic Data Consortium.

Goldman, J. (2011). EasyAlign: An automatic phonetic alignment tool under Praat. In *Proceedings of INTERSPEECH 2011* (pp. 3233–3236). Curran Associates.

Haugen, E. (1972). *First grammatical treatise: An edition, translation and commentary* (2d ed.). Longman.

Herment, S., Tortel, A., Bigi, B., Hirst, D., & Loukina, A. (2014). AixOx, a multi-layered learners' corpus: Automatic annotation. In J. Diaz Pérez & A. Diaz Negrillo (Eds.), *Specialisation and variation in language corpora* (pp. 41–76). Peter Lang.

Heselwood, B. (2019). Phonetic transcription. In J. S. Damico & M. J. Ball (Eds.), *The SAGE encyclopedia of human communication sciences and disorders* (pp. 1359–1362). SAGE.

Hillenbrand, J. M., & Houde, R. A. (1996). Role of F0 and amplitude in the perception of intervocalic glottal stops. *Journal of Speech & Hearing Research*, *39*, 1182–1190.

Hirst, D. (2011). The analysis by synthesis of speech melody: From data to models *<http://www.journalofspeechsciences.org>*. *Journal of Speech Sciences*, *1*(1), 55–83.

Howard, S., & Heselwood, B. (2011). Instrumental and perceptual phonetic analyses: The case for two-tier transcriptions <https://www.tandfonline.com/doi/full/10.3109/02699206.2011.616641?scroll=top&needAccess=true>. *Clinical Linguistics & Phonetics*, *25*, 940–948.

Hualde, J. I., & Prieto, P. (2016). Towards an International Prosodic Alphabet (IPrA) <http://doi.org/10.5334/labphon.11>. *Laboratory Phonology*, *7*(1), 25.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P., & LeBeau, M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. In *Proceedings of INTERSPEECH 2010* (pp. 1914–1917). Curran Associates.

International Phonetic Association (IPA). (1999). *The handbook of the International Phonetic Association*. Cambridge University Press.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation analysis: Studies from the first generation* (pp. 13–31). John Benjamins.

Jespersen, O. (1889). *The articulations of speech sounds*. N. G. Elwert.

Jones, D. (1972). *An outline of English phonetics* (9th ed.). Cambridge University Press.

King, R. (1996). Korean writing. In P. T. Daniels & W. J. Bright (Eds.), *The world's writing systems* (pp. 218–227). Oxford University Press.

Ladefoged, P. (1990). Some reflections on the IPA. *Journal of Phonetics*, *18*, 335–346.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press.

Lee, H. B. (1999). Korean. In *Handbook of the International Phonetic Association* (pp. 120–123). Cambridge University Press.

Lepsius, R. (1863). *Standard alphabet for reducing unwritten languages and foreign graphic systems to a uniform orthography in European letters*. Williams & Norgate.

Liang, M., Lyu, R., & Chiang, Y. (2008). Data-driven approaches to phonetic transcription with integration of automatic speech recognition and grapheme-to-phoneme for spoken Buddhist Sutra. *Computational Linguistics and Chinese Language Processing*, *13*(2), 233–254.

Odden, D. (2005). *Introducing phonology.* Cambridge University Press.

Oller, D. K. (2000). *The emergence of the speech capacity*. Lawrence Erlbaum.

Passy, P. (1907). Alphabet organique. *Le Maître Phonétique*, *22*, 55–57.

Pike, K. L. (1943). *Phonetics*. University of Michigan Press.

Prince, A. S. (1983). Relating to the grid. *Linguistic Inquiry*, *14*, 19–100.

Ramsey, A., Alsharhan, I., & Ahmed, H. (2014). Generation of a phonetic transcription for modern standard Arabic: A knowledge-based model. *Computer Speech and Language*, *28,* 959–978.

Roach, P., Knowles, G., Varadi, T., & Arnfield, S. (1993). MARSEC: A machine-readable spoken English corpus. *Journal of the International Phonetic Association*, *23*(2), 47–53.

Salmon, V. (1972). *The works of Francis Lodwick: A study of his writings in the intellectual context of the seventeenth century*. Longman.

Schillingmann, L., Ernst, J., Keite, V., Wrede, B., Meyer, A. S. & Belke, E. (2018). AlignTool: The automatic temporal alignment of spoken utterances in German, Dutch, and British English for psycholinguistic purposes *<https://doi.org/10.3758/s13428-017-1002-7>*. *Behavior Research Methods*, *50*, 466–489.

Shriberg, L. D., Kwiatkowski, J., & Hoffmann, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech & Hearing Research*, *27*, 456–465.

Skrelin, P., Volskaya, N., Kocharov, D., Evgrafova, K., Glotova, O., & Evdokimova, V. (2010). A fully annotated corpus of Russian speech. In P. Sojka, I. Kopeček, & K. Pala (Eds.), *Proceedings of Text, Speech and Dialogue* (pp. 392–399). Springer.

Sweet, H. (1877). *A handbook of phonetics*. Clarendon Press.

Sweet, H. (1906). *A primer in phonetics* (3d ed.). Clarendon Press.

Taylor, L. J., & Knowles, G. (1988). *Manual of information to accompany the SEC Corpus: The machine readable corpus of spoken English* *<http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM>*.

Tench, P. (1978). On introducing parametric phonetics. *Journal of the International Phonetic Association*, *8*, 34–46.

Van Bael, C., Boves, L., van den Heuval, H., & Strik, H. (2007). Automatic phonetic transcription of large speech corpora *<http://doi.org/10.1016/j.csl.2007.03.003>*. *Computer Speech and Language*, *21*, 652–668.

Walker, G. (2013). Phonetics and prosody in conversation. In J. Sidnell & T. Stivers (Eds.), *Handbook of conversation analysis* (pp. 455–474). Wiley-Blackwell.

Wells, J. C. (2006). *English intonation*. Cambridge University Press.

Wilkins, J. (1668). *An essay towards a real character and a philosophical language*. Sa. Gellibrand.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2017). *The Microsoft 2017 conversational speech recognition system* [Microsoft AI and Research Technical Report MSR-TR-2017-39].

Zue, V. W., & Seneff, S. (1996). Transcription and alignment of the TIMIT database. *Behavior Research Methods*, *50*(2), 466–489.

Related Articles

Contrastive Specification in Phonology

Articulatory Phonetics

Accent in Japanese Phonology

Speech Perception in Phonetics

The Phonetics of Babbling

Direct Perception of Speech

Phonetics of Vowels

Phonetics of Consonants

Audiovisual Speech Perception and the McGurk Effect

Phonetics

Articulatory Phonology