1

Using Speech Corpora in Phonetics Research

1.1 The Place of Corpora in the Phonetic Analysis of Speech

One of the main concerns in phonetic analysis is to find out how speech sounds are transmitted between a speaker and a listener in human speech communication. A speech corpus is a collection of one or more digitized utterances usually containing acoustic data and often marked for annotations. The task in this book is to discuss some of the ways that a corpus can be analyzed to test hypotheses about how speech sounds are communicated. But why is a speech corpus needed for this at all? Why not instead listen to speech, transcribe it, and use the transcription as the main basis for an investigation into the nature of spoken language communication? There is no doubt, as Ladefoged (1995) has explained in his discussion of instrumentation in fieldwork, that being able to hear and re-produce the sounds of a language is a crucial first step in almost any kind of phonetic analysis. Indeed many hypotheses about the way that sounds are used in speech communication stem in the first instance from just this kind of careful listening to speech. However, an auditory transcription is at best an essential initial *hypothesis* – never an objective measure.

The lack of objectivity is readily apparent in comparing the transcriptions of the same speech material across a number of trained transcribers: even when the task is to carry out a fairly broad transcription and with the aid of a speech waveform and spectrogram, there will still be inconsistencies from one transcriber to the next; and all these issues will be considerably aggravated if phonetic detail is to be included in narrower transcriptions or if, as in much fieldwork, auditory phonetic analyses are made of a language with which transcribers are not very familiar. A speech signal on the other hand is a record that does not change: it is, then, the data against which theories can be tested. Another difficulty with building a theory of speech communication on an auditory symbolic transcription of speech is that there are so many ways in which a speech signal is at odds with a segmentation into symbols: there are often no clear boundaries in a speech signal corresponding to the divisions between a string of symbols, and least of all where a layperson might expect to find them, between words.

But, apart from these issues, a transcription of speech can never get to the heart of how the vocal organs, acoustic signal, and hearing apparatus are used to transmit simultaneously many different kinds of information between a speaker and hearer. Consider that the production of /t/ in an utterance tells the listener so much more than "here is a /t/ sound." If the spectrum of the /t/ also has a concentration of energy at a low frequency, then this could be a cue that the following vowel is rounded. At the same time, the alveolar release might provide the listener with information about whether /t/ begins or ends a syllable, a word, or a more major prosodic phrase and whether the syllable is stressed or not. The /t/ might also convey sociophonetic information about the speaker's dialect and quite possibly age group and socioeconomic status (Docherty 2007; Docherty & Foulkes 2005). The combination of /t/ and the following vowel could tell the listener whether the word is prosodically accented and also even say something about the speaker's emotional state.

Understanding how these separate strands of information are interwoven in the details of speech production and the acoustic signal can be accomplished neither by just transcribing speech, nor by analyses of recordings of *individual* utterances. The problem with analyses of individual utterances is that they risk being idiosyncratic: this is not only because of all the different ways that speech can vary according to context, but also because the anatomical and speaking-style differences between speakers all leave their mark on the acoustic signal: therefore, an analysis of a handful of speech sounds in one or two utterances may give a distorted presentation of the general principles according to which speech communication takes place.

The issues raised above and the need for speech corpora in phonetic analysis in general can be considered from the point of view of other more recent theoretical developments: that the relationship between phonemes and speech is *stochastic*. This is an important argument that has been put forward by Janet Pierrehumbert in a number of papers in recent years (e.g., 2002, 2003a, 2003b, 2006). On the one hand there are almost certainly different levels of abstraction or, in terms of the episodic/exemplar models of speech perception and production developed by Pierrehumbert and others (Bybee 2001; Goldinger 1998, 2000; Johnson 1997), *generalizations* that allow native speakers of a language to recognize that *tip* and *pit* are composed of the same three sounds but in the opposite order. Now it is also undeniable that different languages, and certainly different varieties of the same language, often make broadly similar sets of phonemic contrasts: thus in many languages, differences of meaning are established as a result of contrasts between voiced and voiceless stops, or between oral stops and nasal stops at the same place of articulation, or between rounded and unrounded vowels of the same height, and so on. But what has never been demonstrated is that two languages that make similar sets of contrasts do so phonetically in exactly the same way. These differences might be subtle, but they are nevertheless present, which means that such differences must have been learned by the speakers of the language or community.

But how do such differences arise? One way in which they are unlikely to be brought about is because languages or their varieties choose their sound systems from a finite set of universal features. At least so far, no one has been able to demonstrate that the number of possible permutations that could be derived even from the most comprehensive of articulatory or auditory feature systems could account for the myriad

of ways that the sounds of dialects and languages do in fact differ. It seems instead that, although the sounds of languages undeniably conform to consistent patterns (as demonstrated in the ground-breaking study of vowel dispersion by Liljencrants & Lindblom 1972), there is also an arbitrary, stochastic component to the way in which the association between abstractions like phonemes and features evolves and is learned by children (Beckman et al. 2007; Edwards & Beckman 2008; Munson et al. 2005).

Recently, this stochastic association between speech on the one hand and phonemes on the other has been demonstrated computationally using so-called agents equipped with simplified vocal tracts and hearing systems who imitate each other over a large number of computational cycles (Wedel 2006, 2007). The general conclusion from these studies is that while stable phonemic systems emerge from these initially random imitations, there are a potentially infinite number of different ways in which phonemic stability can be achieved (and then shifted in sound change – see also Boersma & Hamann 2008). A very important idea to emerge from these studies is that the phonemic stability of a language does not require *a priori* a selection to be made from a pre-defined universal feature system, but might emerge instead as a result of speakers and listeners copying each other imperfectly (Oudeyer 2002, 2004).

If we accept the argument that the association between phonemes and the speech signal is not derived deterministically by making a selection from a universal feature system, but is instead arrived at stochastically by learning generalizations across produced and perceived speech data, then it necessarily follows that analyzing corpora of speech must be one of the important ways in which we can understand how different levels of abstraction such as phonemes and other prosodic units are communicated in speech.

Irrespective of these theoretical issues, speech corpora have become increasingly important since the 1980s as the primary material on which to train and test human–machine communication systems. Some of the same corpora that have been used for technological applications have also formed part of basic speech research (see 1.2 for a summary of these). One of the major benefits of these corpora is that they foster a much needed interdisciplinary approach to speech analysis, as researchers from different disciplinary backgrounds apply and exchange a wide range of techniques for analyzing the data.

Corpora that are suitable for phonetic analysis may become available with the increasing need for speech technology systems to be trained on various kinds of fine phonetic detail (Carlson & Hawkins 2007). It is also likely that corpora will be increasingly useful for the study of sound change as more archived speech data becomes available with the passage of time, allowing sound change to be analyzed either longitudinally in individuals (Harrington 2006; Labov & Auger 1998) or within a community using so-called real-time studies (for example, by comparing the speech characteristics of subjects from a particular age group recorded today with those of a comparable age group and community recorded several years ago – see Sankoff 2005; Trudgill 1988). Nevertheless, most types of phonetic analysis still require collecting small corpora that are dedicated to resolving a particular research question and associated hypotheses; some of the issues in designing such corpora are discussed in 1.3.

Finally, before covering some of these design criteria, it should be pointed out that speech corpora are by no means necessary for every kind of phonetic investigation, and indeed many of the most important scientific breakthroughs in phonetics in the last 50 years have taken place without analyses of large speech corpora. For example, speech corpora are usually not needed for various kinds of articulatory-to-acoustic modeling nor for many kinds of studies in speech perception in which the aim is to work out, often using speech synthesis techniques, the sets of cues that are functional – that is, relevant for phonemic contrasts.

1.2 Existing Speech Corpora for Phonetic Analysis

The need to provide an increasing amount of training and testing material has been one of the main driving forces in creating speech and language corpora in recent years. Various sites for their distribution have been established and some of the more major ones include: the Linguistic Data Consortium (Reed et al. 2008),[1] which is a distribution site for speech and language resources and is located at the University of Pennsylvania; ELRA,[2] the European Language Resources Association, established in 1995, which validates, manages, and distributes speech corpora and whose operational body is ELDA[3] (Evaluations and Language resources Distribution Agency). There are also a number of other repositories for speech and language corpora, including the Bavarian Archive for Speech Signals[4] at the University of Munich, various corpora at the Center for Spoken Language Understanding at the University of Oregon,[5] the TalkBank consortium at Carnegie Mellon University,[6] and the DOBES archive of endangered languages at the Max Planck Institute in Nijmegen.[7]

Most of the corpora from these organizations serve primarily the needs for speech and language technology, but there are a few large-scale corpora that have also been used to address issues in phonetic analysis, including the Switchboard and TIMIT corpora of American English. The Switchboard corpus (Godfrey et al. 1992) includes over 600 telephone conversations from 750 adult American English speakers of a wide range of ages and varieties from both genders and was recently analyzed by Bell et al. (2003) in a study investigating the relationship between predictability and the phonetic reduction of function words. The TIMIT database (Garofolo et al. 1993; Lamel et al. 1986) has been one of the most studied corpora for assessing the performance of speech-recognition systems since the 1980s. It includes 630 talkers and 2,342 different read speech sentences, comprising over 5 hours of speech, and has been included in various phonetic studies on topics such as variation between speakers (Byrd 1992), the acoustic characteristics of stops (Byrd 1993), the relationship between gender and dialect (Byrd 1994), word and segment duration (Keating et al. 1994a), vowel and consonant reduction (Manuel et al. 1992), and vowel normalization (Weenink 2001). One of the most extensive corpora of a European language other than English is the Dutch CGN corpus[8] (Oostdijk 2000; Pols 2001). This is the largest corpus of contemporary Dutch spoken by adults in Flanders and the Netherlands and includes around 800 hours of speech. In the last few years, it has been used to study the sociophonetic variation in diphthongs (Jacobi et al. 2007). For German, the Kiel Corpus of Speech[9] includes several hours of speech

annotated at various levels (Simpson 1998; Simpson et al. 1997) and has been instrumental in studying different kinds of connected speech processes (Kohler 2001; Simpson 2001; Wesener 2001).

One of the most successful corpora for studying the relationship between discourse structure, prosody, and intonation has been the HCRC Map Task Corpus[10] (Anderson et al. 1991), containing 18 hours of annotated spontaneous speech recorded from 128 two-person conversations according to a task-specific experimental design (see below for further details). The Australian National Database of Spoken Language[11] (Millar et al. 1994, 1997) also contains a similar range of map-task data for Australian English. These corpora have been used to examine the relationship between speech clarity and the predictability of information (Bard et al. 2000), and also to investigate the way that boundaries between dialogue acts interact with intonation and suprasegmental cues (Stirling et al. 2001). More recently, two corpora have been developed primarily for phonetic and basic speech research: one is the Buckeye Corpus,[12] consisting of 40 hours of spontaneous American English speech annotated at word and phonetic levels (Pitt et al. 2005), which has recently been used to model /t, d/ deletion (Raymond et al. 2006). Another is the Nationwide Speech Project (Clopper & Pisoni 2006), which is especially useful for studying differences in American varieties. It contains 60 speakers from six regional varieties of American English and parts of it are available from the Linguistic Data Consortium.

Databases of speech physiology are much less common than those of speech acoustics, largely because they have not evolved in the context of training and testing speech-technology systems (which is the main source of funding for speech-corpus work). One exception is the ACCOR speech database (Marchal & Hardcastle 1993; Marchal et al. 1991), developed in the 1990s to investigate coarticulatory phenomena in a number of European languages and which includes laryngographic, airflow, and electropalatographic data (the database is available from ELRA). Another is the University of Wisconsin X-ray Microbeam Speech Production Database (Westbury 1994) which includes acoustic and movement data from 26 female and 22 male speakers of a Midwest dialect of American English aged between 18 and 37 years. Thirdly, the MOCHA-TIMIT[13] database (Wrench & Hardcastle 2000) is made up of synchronized movement data from the supralaryngeal articulators, electropalatographic data, and a laryngographic signal of part of the TIMIT database produced by subjects of different English varieties. These databases have been incorporated into phonetic studies in various ways: for example, the Wisconsin database was used by Simpson (2002) to investigate the differences between male and female speech, and the MOCHA-TIMIT database formed part of a study by Kello and Plaut (2003) to explore feedforward learning association between articulation and acoustics in a cognitive speech-production model.

Finally, there are many opportunities to obtain quantities of speech data from archived broadcasts (e.g., in Germany from the Institut für Deutsche Sprache in Mannheim; in the UK from the BBC). These are often acoustically of high quality. However, it is unlikely they will have been annotated, unless they have been incorporated into an existing corpus design, as was the case in the development of the Machine Readable Corpus of Spoken English (MARSEC) created by Roach et al. (1993) based on recordings from the BBC.

1.3 Designing Your Own Corpus

Unfortunately, most kinds of phonetic analysis still require building a speech corpus that is designed to address a specific research question. In fact, existing large-scale corpora of the kind sketched above are very rarely used in basic phonetic research, partly because, no matter how extensive they are, a researcher inevitably finds that one or more aspects of the speech corpus (e.g., speakers, types of materials, speaking styles) are insufficiently covered for the research question to be completed. Another problem is that an existing corpus may not have been annotated in the way that is needed. A further difficulty is that the same set of speakers might be required for a follow-up speech-perception experiment after an acoustic corpus has been analyzed, and inevitably access to the subjects of the original recordings is out of the question, especially if the corpus was created a long time ago.

Assuming that you have to put together your own speech corpus, then various issues in design need to be considered, to make sure not only that the corpus is adequate for answering the specific research questions that are required of it, but also that it is possibly re-usable by other researchers at a later date. It is important to give careful thought to designing the speech corpus, because collecting and especially annotating almost any corpus is usually very time-consuming. Some non-exhaustive issues, based to a certain extent on Schiel and Draxler (2004), are outlined below. The brief review does not cover recording acoustic and articulatory data from endangered languages which brings an additional set of difficulties as far as access to subjects and designing materials are concerned (see in particular Ladefoged 1995, 2003).

1.3.1 Speakers

Choosing the speakers is obviously one of the most important issues in building a speech corpus. Some primary factors to take into account include the distribution of speakers by gender, age, first language, and variety (dialect); it is also important to document any known speech or hearing pathologies. For sociophonetic investigations, or studies specifically concerned with speaker characteristics, a further refinement according to many other factors such as educational background, profession, or socioeconomic group (to the extent that this is not covered by variety) is also likely to be important (see also Beck 2005 for a detailed discussion of the parameters of a speaker's vocal profile based to a large extent on Laver 1980, 1991). All of the above-mentioned primary factors are known to exert quite a considerable influence on the speech signal and therefore have to be controlled for in any experiment comparing two or more speaking groups. Thus it would be inadvisable in comparing, say, speakers of two different varieties to have a predominance of male speakers in one group, and female speakers in another, or one group with mostly young and the other with mostly older speakers. Whatever speakers are chosen, it is, as Schiel and Draxler (2004) comment, of great importance that as many details of the speakers are documented as possible (see also Millar 1991), should the need arise to check subsequently whether the speech data might have been influenced by a particular speaker-specific attribute.

The next most important criterion is the number of speakers. Following Gibbon et al. (1997), speech corpora of between one and five speakers are typical in the

context of speech synthesis development, while more than 50 speakers are needed for adequately training and testing systems for the automatic recognition of speech. For most experiments in experimental phonetics of the kind reported in this book, a speaker sample size within the range of 10 to 20 is usual. In almost all cases, experiments involving invasive techniques such as electromagnetic articulometry and electropalatography (discussed in Chapters 5 and 7) rarely have more than five speakers because of the time taken to record and analyze the speech data and the difficulty in finding subjects.

1.3.2 Materials

An equally important consideration in designing any corpus is the choice of materials. Four of the main parameters discussed in Schiel and Draxler (2004) according to which materials are chosen are *vocabulary*, *phonological distribution*, *domain*, and *task*.

Vocabulary in a speech-technology application such as automatic speech recognition derives from the intended use of the corpus: so a system for recognizing digits must obviously include the digits as part of the training material. In many phonetics experiments, a choice has to be made between real words of the language and non-words. In either case, it will be necessary to control for a number of phonological criteria, some of which are outlined below (see also Rastle et al. 2002 and the associated website[14] for a procedure for selecting non-words according to numerous phonological and lexical criteria). Since both lexical frequency and neighborhood density have been shown to influence speech production (Luce & Pisoni 1998; Wright 2003), then it could be important to control for these factors as well, possibly by retrieving these statistics from a corpus such as Celex (Baayen et al. 1995). Lexical frequency, as its name suggests, is the estimated frequency with which a word occurs in a language: at the very least, confounds between words of very high frequency, such as between function words which tend to be heavily reduced even in read speech, and less frequently occurring content words should be avoided. Words of high neighborhood density can be defined as those for which many other words exist by substituting a single phoneme (e.g., *man* and *van* are neighbors according to this criterion). Neighborhood density is less commonly controlled for in phonetics experiments although, as recent studies have shown (Munson & Solomon 2004; Wright 2003), it too can influence the phonetic characteristics of speech sounds.

The words that an experimenter wishes to investigate in a speech-production experiment should not be presented to the subject in a list (which induces a so-called list prosody in which the subject chunks the lists into phrases, often with a falling melody and phrase-final lengthening on the last word, but a level or rising melody on all the others); these are often displayed on a screen individually or incorporated into a so-called carrier phrase. Both of these conditions will go some way towards neutralizing the effects of sentence-level prosody, i.e., towards ensuring that the intonation, phrasing, rhythm, and accentual pattern are the same from one target word to the next. Sometimes filler words need to be included in the list, in order to draw the subject's attention away from the design of the experiment. This is important because, if any parts of the stimuli become predictable, then a subject might well reduce them phonetically, given the relationship between redundancy and predictability (Fowler & Housum 1987; Hunnicutt 1985; Lieberman 1963).

For some speech-technology applications, the materials are specified in terms of their *phonological distribution*. For almost all studies in experimental phonetics, controlling for the phonological composition of the target words, in terms of factors such as their lexical-stress pattern, number of syllables, syllable composition, and segmental context, is essential because these all exert an influence on the utterance. In investigations of prosody, materials are sometimes constructed in order to elicit certain kinds of phrasing, accentual patterns, or even intonational melodies. In Silverman and Pierrehumbert (1990), two subjects produced a variety of phrases like *Ma Le Mann*, *Ma Lemm*, and *Mamalie Lemonick* with a prosodically accented initial syllable and identical intonation melody: they used these materials in order to investigate whether the timing of the pitch-accent was dependent on factors such as the number of syllables in the phrase and the presence or absence of word boundaries. In various experiments by Keating and colleagues (e.g., Keating et al. 2003), French, Korean, and Taiwanese subjects produced sentences that had been constructed to control for different degrees of boundary strength. Thus their French materials included sentences in which /na/ occurred at the beginning of phrases at different positions in the prosodic hierarchy, such as initially in the accentual phrase (*Tonton, Tata, **Na**dia et Paul arriveront demain*) and syllable-initially (*Tonton et **Ana**belle . . .*). In Harrington et al. (2000), materials were designed to elicit the contrast between accented and deaccented words. For example, the name *Beaber* was accented in the introductory statement *This is Hector **Beaber***, but deaccented in the question *Do you want **Anna** Beaber or **Clara** Beaber* (in which the nuclear accents falls on the preceding first name). Creating corpora such as these can be immensely difficult, however, because there will always be some subjects who do not produce words as the experimenter wishes (for example by not fully deaccenting the target words in the last example) or, if they do, they might introduce unwanted variations in other prosodic variables. The general point is that subjects usually need to have some training in the production of materials in order to produce them with the degree of consistency required by the experimenter. However, this leads to the additional concern that the productions might not really be representative of prosody produced in spontaneous speech by the wider population.

These are some of the reasons why the production of prosody is sometimes studied using map-task corpora (Anderson et al. 1991) of the kind referred to earlier, in which a particular prosodic pattern is not prescribed, but instead emerges more naturally out of a dialogue or situational context. The map task is an example of a corpus that falls into the category defined by Schiel and Draxler (2004) of being *restricted by domain*. In the map task, two dialogue partners are given slightly different versions of the same map and one has to explain to the other how to navigate a route between two or more points along the map. An interesting variation on this is due to Peters (2006) in which the dialogue partners discuss the contents of two slightly different video recordings of a popular soap opera that both subjects happen to be interested in: the interest factor has the potential additional advantage that the speakers will be distracted by the content of the task, and thereby produce speech in a more natural way. In either case, a fair degree of prosodic variation and spontaneous speech are guaranteed. At the same time, the speakers' choice of prosodic patterns and lexical items tends to be reasonably constrained, allowing comparisons between different speakers on this task to be made in a meaningful way.

In some types of corpora, a speaker will be instructed to solve a particular *task*. The instructions might be fairly general, as in the map task or the video scenario described above, or they might be more specific, such as describing a picture or answering a set of questions. An example of a task-specific recording is in Schafer et al. (2000) who used a cooperative game task in which subjects disambiguated in their productions ambiguous sentences such as *move the square with the triangle* (meaning either: move a house-like shape consisting of a square with a triangle on top of it; or: move a square piece with a separate triangular piece). Such a task allows experimenters to restrict the dialogue to a small number of words, it distracts speakers from the task at hand (since speakers have to concentrate on how to move pieces rather than on what they are saying) while at the same time eliciting precisely the different kinds of prosodic parsings required by the experimenter in the same sequence of words.

1.3.3 Some further issues in experimental design

Experimental design in the context of phonetics is to do with making choices about the speakers, materials, number of repetitions, and other issues that form part of the experiment in such a way that the validity of a hypothesis can be quantified and tested statistically. The summary below touches only very briefly on some of the matters to be considered at the stage of laying out the experimental design, and the reader is referred to Robson (1994), Shearer (1995), and Trochim (2007) for many further useful details. What is presented here is also mostly about some of the design criteria that are relevant for the kind of experiment leading to a statistical test such as analysis of variance (ANOVA). It is quite common for ANOVAs to be applied to experimental speech data, but this is obviously far from the only kind of statistical test that phoneticians need to apply, so some of the issues discussed will not necessarily be relevant for some types of phonetic investigation.

In a certain kind of experiment that is common in experimental psychology and experimental phonetics, a researcher will often want to establish whether a **dependent** variable is affected by one or more **independent** variables. The dependent variable is what is measured – for the kind of speech research discussed in this book, the dependent variable might be any one of duration, a formant frequency at a particular time point, the vertical or horizontal position of the tongue at a displacement maximum, and so on. These are all examples of **continuous** dependent variables because, like age or temperature, they can take on an infinite number of possible values within a certain range. Sometimes the dependent variable might be **categorical**, as in eliciting responses from subjects in speech-perception experiments in which the response is a specific category (e.g., a listener labels a stimulus as either /ba/ or /pa/). Categorical variables are common in sociophonetic research in which counts are made of data (e.g., a count of the number of times that a speaker produces /t/ with or without glottalization).

The independent variable, or factor, is what you believe has an influence on the dependent variable. One type of independent variable that is common in experimental phonetics comes about when a comparison is made between two or more groups of speakers, such as between male and female speakers. This type of independent variable is sometimes (for obvious reasons) called a **between-speaker factor**, which in

this example might be given a name like **Gender**. Some further useful terminology is to do with the **number of levels of the factor**. For this example, **Gender** has two levels, **male** and **female**. The same speakers could of course also be coded for other between-speaker factors. For example, the same speakers might be coded for a factor **Variety** with three levels: **Standard English**, **Estuary English**, and **Cockney**. **Gender** and **Variety** in this example are **nominal** because the levels are not rank ordered in any way. If the ordering matters, then the factor is **ordinal** (for example **Age** could be an ordinal factor if you wanted to assess the effects on increasing age of the speakers).

Each speaker that is analyzed can be assigned just one level of each between-speaker factor: so each speaker will be coded as either **male** or **female**, and as **Standard English**, **Estuary English**, or **Cockney**. This example would also sometimes be called a 2×3 design, because there are two factors with two (**Gender**) and three (**Variety**) levels. An example of a $2 \times 3 \times 2$ design would have three factors with the corresponding number of levels, e.g., the subjects are coded not only for **Gender** and **Variety** as before, but also for **Age** with two levels, **young** and **old**. Some statistical tests require that the design should be approximately **balanced**: specifically, a given between-subjects factor should have equal numbers of subjects distributed across its levels. For the previous example with two factors, **Gender** and **Variety**, a balanced design would be one that had 12 speakers, 6 males and 6 females, and 2 male and 2 female speakers per variety. Another consideration is that the more between-subjects factors you include, then evidently the greater the number of speakers from which recordings have to be made. Experiments in phonetics are often restricted to no more than two or three between-speaker factors, not just because of considerations of the size of the subject pool, but also because the statistical analysis in terms of interactions becomes increasingly unwieldy for a larger number of factors.

Now suppose you wish to assess whether these subjects show differences of vowel duration in words with a final /t/ like *white* compared with words with a final /d/ like *wide*. In this case, the design might include a factor **Voice** and it has two levels: **[-voice]** (words like *white*) and **[+voice]** (words like *wide*). One of the things that make this type of factor very different from the between-speaker factors considered earlier is that subjects produce (i.e., are measured on) all of the factor's levels – that is, the subjects will produce words that are both **[-voice] and [+voice]**. **Voice** in this example would sometimes be called a within-subject or **within-speaker factor** and because subjects are measured on all of the levels of **Voice**, it is also said to be **repeated**. This is also the reason why, if you wanted to use an ANOVA to work out whether **[+voice]** and **[-voice]** words differed in vowel duration, and also whether such a difference manifested itself in the various speaker groups, you would have to use a **repeated-measures ANOVA**. Of course, if one group of subjects produced the **[-voice]** words and another group the **[+voice]** words, then **Voice** would not be a repeated factor and so a conventional ANOVA could be applied. However, in experimental phonetics this would not be a sensible approach, not just because you would need many more speakers, but also because the difference between **[-voice]** and **[+voice]** words in the dependent variable (vowel duration) would then be confounded with speaker differences. So this is why repeated or within-speaker factors are very common in

experimental phonetics. Of course, in the same way that there can be more than one between-speaker factor, there can also be two or more within-speaker factors. For example, if the **[-voice]** and **[+voice]** words were each produced at a slow and a fast rate, then **Rate** would also be a within-speaker factor with two levels (**slow** and **fast**). **Rate**, like **Voice**, is a within-speaker factor because the same subjects have been measured once at a slow, and once at a fast rate.

The need to use a repeated-measures ANOVA comes about, then, because the subject is measured on all the levels of a factor and (somewhat confusingly) it has nothing whatsoever to do with **repeating the same level of a factor in speech production**, which in experimental phonetics is rather common. For example, the subjects might be asked to repeat (in some randomized design) *white* at a slow rate five times. This repetition is done to counteract the inherent variation in speech production. One of the very few uncontroversial facts of speech production is that no subject can produce the same utterance twice, even under identical recording conditions, in exactly the same way. So since a single production of a target word could just happen to be a statistical aberration, researchers in experimental phonetics usually have subjects produce exactly the same materials many times over: this is especially so in physiological studies, in which this type of inherent token-to-token variation is usually so much greater in articulatory than in acoustic data. However, it is important to remember that repetitions of the same level of a factor (the multiple values from each subject's slow production of *white*) cannot be entered into many standard statistical tests such as a repeated-measures ANOVA and so they typically need to be averaged (see Max & Onghena 1999 for some helpful details on this). So even if, as in the earlier example, a subject repeats *white* and *wide* each several times at both slow and fast rates, only four values per subject can be entered into the repeated-measures ANOVA (i.e., the four mean values for each subject of: *white* at a slow rate, *white* at a fast rate, *wide* at a slow rate, *wide* at a fast rate). Consequently, the number of repetitions of identical materials should be kept sufficiently low because otherwise a lot of time will be spent recording and annotating a corpus without really increasing the likelihood of a significant result (on the assumption that the values that are entered into a repeated-measures ANOVA averaged across 10 repetitions of the same materials may not differ a great deal from the averages calculated from 100 repetitions produced by the same subject). The number of repetitions and indeed total number of items in the materials should in any case be kept within reasonable limits because otherwise subjects are likely to become bored and, especially in the case of physiological experiments, fatigued, and these types of paralinguistic effects may well in turn influence their speech production.

The need to average across repetitions of the same materials for certain kinds of statistical test described in Max and Onghena (1999) justifiably seems bizarre to many experimental phoneticians, especially in speech physiology research in which the variation, even in repeating the same materials, may be so large that an average or median becomes fairly meaningless. Fortunately, there have recently been considerable advances in the statistics of **mixed-effects modeling** (see the special edition by Forster & Masson 2008 on emerging data analysis and various papers within that; see also Baayen 2008), which provides an alternative to the classical use of a repeated-measures ANOVA. One of the many advantages of this technique is that there is no need to average across repetitions (Quené & van den Bergh 2008). Another is that

it provides a solution to the so-called language-as-fixed-effect problem (Clark 1973). The full details of this matter need not detain us here: the general concern raised in Clark's (1973) influential paper is that, in order to be sure that the statistical results generalize not only beyond the subjects of your experiment but also beyond the language materials (i.e., are not just specific to *white*, *wide*, and the other items of the word list), two separate (repeated-measures) ANOVAs need to be carried out, one so-called by-subjects and the other by-items (see Johnson 2008 for a detailed exposition using speech data in R). The output of these two tests can then be combined using a formula to compute the joint *F*-ratio (and therefore the significance) from both of them. By contrast, there is no need in mixed-effects modeling to carry out and to combine two separate statistical tests in this way: instead, the subjects and the words can be entered as so-called random factors into the same calculation.

Since much of the cutting-edge mixed-effects modeling research in statistics has been carried out in R in the last 10 years, there are corresponding R functions for carrying out mixed-effects modeling that can be directly applied to speech data, without the need to go through the often very tiresome complications of exporting the data, sometimes involving rearranging rows and columns for analysis using the more traditional commercial statistical packages.

1.3.4 Speaking style

A wide body of research since 1960 has shown that speaking style influences speech production characteristics: in particular, the extent of coarticulatory overlap, vowel centralization, consonant lenition and deletion are all likely to increase in progressing from citation-form speech, in which words are produced in isolation or in a carrier phrase, to read speech and to fully spontaneous speech (Moon & Lindblom 1994). In some experiments, speakers are asked to produce speech at different rates so that the effect of increasing or decreasing tempo on consonants and vowels can be studied. However, in the same way that it can be problematic to get subjects to produce controlled prosodic materials consistently (see 1.3.2), the task of making subjects vary speaking rate is not without its difficulties. Some speakers may not vary their rate a great deal in changing from "slow" to "fast" and one person's slow speech may be similar to another subject's fast rate. Subjects may also vary other prosodic attributes in switching from a slow to a fast rate. In reading a target word within a carrier phrase, subjects may well vary the rate of the carrier phrase but not the focused target word that is the primary concern of the investigation: this might happen if the subject (not unjustifiably) believes the target word to be communicatively the most important part of the phrase, as a result of which it is produced slowly and carefully at all rates of speech.

The effect of emotion on prosody is a very much under-researched area that also has important technological applications in speech-synthesis development. However, eliciting different kinds of emotion, such as a happy or sad speaking style, is problematic. It is especially difficult, if not impossible, to elicit different emotional responses to the same read material, and, as Campbell (2002) notes, subjects often become self-conscious and suppress their emotions in an experimental task. An alternative then might be to construct passages that describe scenes associated with different emotional content, but then even if the subject achieves a reasonable degree

of variation in emotion, any influence of emotion on the speech signal is likely to be confounded with the potentially far greater variation induced by factors such as the change in focus and prosodic accent, the effects of phrase-final lengthening, and the use of different vocabulary. (There is also the independent difficulty of quantifying the extent of happiness and sadness with which the materials were produced.) Another possibility is to have a trained actor produce the same materials in different emotional speaking styles (e.g., Pereira 2000), but whether this type of forced variation by an actor really carries over to emotional variation in everyday communication can only be assumed and is not easily verified (however see, e.g., Campbell 2002, 2004; and Douglas-Cowie et al. 2003 for some recent progress in approaches to creating corpora for "emotion" and expressive speech).

1.3.5 Recording setup[15]

Many experiments in phonetics are carried out in a sound-treated recording studio in which the effects of background noise can be largely eliminated and in which the speaker is seated at a controlled distance from a high-quality microphone. Since, with the possible exception of some fricatives, most of the phonetic content of the speech signal is contained below 8 kHz, and taking into account the Nyquist theorem (see also Chapter 8) that only frequencies below half the sampling frequency can be faithfully reproduced digitally, the sampling frequency is typically at least 16 kHz in recording speech data. The signal should be recorded in an uncompressed or PCM (pulse code modulation) format and the amplitude of the signal is typically quantized in 16 bits: this means that the amplitude of each sampled data value occurs at one of a number of 2^{16} discrete steps, which is usually considered adequate for representing speech digitally. With the introduction of the audio CD standard, a sampling frequency of 44.1 kHz and its divider 22.05 kHz are also common. An important consideration in any recording of speech is to set the input level correctly: if it is too high, a distortion known as clipping can result, while if it is too low, then the amplitude resolution will also be too low. For some types of investigations of communicative interaction between two or more speakers, it is possible to make use of a stereo microphone as a result of which data from the separate channels are interleaved or multiplexed (in which the samples from, e.g., the left and right channels are contained in alternating sequence). However, Schiel and Draxler (2004) recommend instead using separate microphones since interleaved signals may be more difficult to process in some signal processing systems – for example, at the time of writing, the speech signal processing routines in Emu cannot be applied to stereo signals.

There are a number of file formats for storing digitized speech data including a raw format which has no header and contains only the digitized signal; NIST SPHERE defined by the National Institute for Standards and Technology, USA consisting of a readable header in plain text (7 bit US ASCII) followed by the signal data in binary form; and most commonly the WAVE file format which is a subset of Microsoft's RIFF specification for the storage of multimedia files.

If you make recordings beyond the recording studio, and in particular if this is done without technical assistance, then, apart from the sampling frequency and bit-rate, factors such as background noise and the distance of the speaker from the microphone need to be very carefully monitored. Background noise may be

especially challenging: if you are recording in what seems to be a quiet room, it is nevertheless important to check that there is no other hum or interference from other electrical equipment such as an air-conditioning unit. Although present-day personal and notebook computers are equipped with built-in hardware for playing and recording high-quality audio signals, Draxler (2008) recommends using an external device such as a USB headset for recording speech data. The recording should only be made onto a laptop in battery mode, because the AC power source can sometimes introduce noise into the signal.[16]

One of the difficulties with recording in the field is that you usually need separate pieces of software for recording the speech data and for displaying any prompts and recording materials to the speaker. Recently, Draxler and Jänsch (2004) have provided a solution to this problem by developing a freely available, platform-independent software system for handling multi-channel audio recordings known as SpeechRecorder.[17] It can record from any number of audio channels and has two screens that are seen separately by the subject and by the experimenter. The first of these includes instructions about when to speak as well as the script to be recorded. It is also possible to present auditory or visual stimuli instead of text. The screen for the experimenter provides information about the recording level, details of the utterance to be recorded, and which utterance number is being recorded. One of the major advantages of this system is not only that it can be run from almost any PC, but also that the recording sessions can be done with this software over the internet. In fact, SpeechRecorder has recently been used just for this purpose (Draxler & Jänsch 2007) in the collection of data from teenagers in a very large number of schools from all around Germany. It would have been very costly to have to travel to the schools, so being able to record and monitor the data over the internet was an appropriate solution in this case. This type of internet solution would be even more useful if speech data were needed across a much wider geographical area.

The above is a description of procedures for recording acoustic speech signals (see also Draxler 2008 for further details), but it can to a certain extent be applied to the collection of physiological speech data. There is articulatory equipment for recording aerodynamic, laryngeal, and supralaryngeal activity and some information from lip movement could even be obtained with video recordings synchronized with the acoustic signal. However, video information is rarely precise enough for most forms of phonetic analyses. Collecting articulatory data is inherently complicated because most of the vocal organs are hidden and so the techniques are often invasive (see various chapters in Hardcastle & Hewlett 1999 and Harrington & Tabain 2004 for a discussion of some of these articulatory techniques). A physiological technique such as electromagnetic articulometry described in Chapter 5 also requires careful calibration; and physiological instrumentation tends to be expensive, restricted to laboratory use, and generally not easily useable without technical assistance. The variation within and between subjects in physiological data can be considerable, often requiring an analysis and statistical evaluation subject by subject. The synchronization of the articulatory data with the acoustic signal is not always a trivial matter and analyzing articulatory data can be very time-consuming, especially if data are recorded from several articulators. For all these reasons, there are far fewer experiments in phonetics using articulatory techniques than there are using acoustic techniques. At the same time, physiological techniques can provide insights into

speech-production control and timing which cannot be accurately inferred from acoustic techniques alone.

1.3.6 Annotation

The annotation of a speech corpus refers to the creation of symbolic information that is related to the signals of the corpus in some way. It is always necessary for annotations to be time-aligned with the speech signals: for example, there might be an orthographic transcript of the recording and then the words might be further tagged for syntactic category, or sentences for dialogue acts, with these annotations being assigned any markers to relate them to the speech signal in time. In the phonetic analysis of speech, the corpus usually has to be **segmented and labeled**, which means that symbols are linked to the physical time scale of one or more signals. As described more fully in Chapter 4, a symbol may be either a segment that has a certain duration or else an event that is defined by a single point in time. The segmentation and labeling are often done manually by an expert transcriber with the aid of a spectrogram. Once part of the database has been manually annotated, then it can sometimes be used as training material for the automatic annotation of the remainder. The Institute of Phonetics and Speech Processing of the University of Munich makes extensive use of the Munich automatic segmentation system (MAUS) developed by Schiel (1999, 2004) for this purpose. MAUS typically requires a segmentation of the utterance in words, based on which statistically weighted hypotheses of sub-word segments can be calculated and then verified against the speech signal. Exactly this procedure was used to provide an initial phonetic segmentation of the acoustic signal for the corpus of movement data discussed in Chapter 5.

Manual segmentation tends to be more accurate than automatic segmentation and it has the advantage that segmentation boundaries can be perceptually validated by expert transcribers (Gibbon et al. 1997): certainly, it is always necessary to check the annotations and segment boundaries established by an automatic procedure before any phonetic analysis can take place. However, an automatic procedure has the advantage over manual procedures not only of complete acoustic consistency but especially that annotation is accomplished much more quickly.

One of the reasons why manual annotation is complicated is because of the continuous nature of speech: it is very difficult to make use of external acoustic evidence to place a segment boundary between the consonants and vowel in a word like *wheel* because the movement between them is not discrete but continuous. Another major source of difficulty in annotating continuous or spontaneous speech is that there will be frequent mismatches between the phonetic content of the signal and the citation-form pronunciation. Thus *run past* might be produced with assimilation and deletion as [ɹʌmpɑːs], *actually* as [aʃli] and so on (Laver 1994). One of the difficulties for a transcriber is in deciding upon the extent to which reduction has taken place and whether segments overlap completely or partially. Another is in aligning the reduced forms with citation-form dictionary entries, which is sometimes done in order to measure subsequently the extent to which segmental reduction has taken place in different contexts (see Harrington et al. 1993 and Appendix B of the website related to this book for an example of a matching algorithm to link reduced and citation forms, and Johnson 2004b for a technique which, like

Harrington et al. 1993, is based on dynamic programming for aligning the two types of transcription).

The inherent difficulty in segmentation can be offset to a certain extent by following some basic procedures in carrying out this task. One fairly obvious one is that it is best not to segment and label any more of the corpus than is necessary for addressing the hypotheses that are to be solved in analyzing the data phonetically, given the amount of time that manual segmentation and labeling take. A related point (which is discussed in further detail in Chapter 4) is that the database needs to be annotated in such a way that the speech data that is required for the analysis can be queried or extracted without too much difficulty. One way to think about manual annotation in phonetic analysis is that it acts as a form of scaffolding (which may not form part of the final analysis) allowing a user to access the data of interest. But, just like scaffolding, the annotation needs to be firmly grounded, which means that segment boundaries should be placed at relatively unambiguous acoustic landmarks if at all possible. For example, if you are interested in the rate of transition between semi-vowels and vowels in words like *wheel*, then it is probably not a good idea to have transcribers try to find the boundary at the juncture between the consonants and vowel for the reasons stated earlier that it is very difficult to do so, based on any objective criteria (leading to the additional problem that the consistency between separate transcribers might not be very high). Instead, the words might be placed in a carrier phrase so that the word onset and offset can be manually marked: the interval between the word boundaries could then be analyzed algorithmically based on objective acoustic factors such as the maximum rate of formant change.

For all the reasons discussed so far, there should never really be any need for a complete, exhaustive segmentation and labeling of entire utterances into phonetic segments: it is too time-consuming, unreliable, and is probably in any case not necessary for most types of phonetic analyses. If this type of exhaustive segmentation really is needed, as perhaps in measuring the variation in the duration of vowels and consonants in certain kinds of studies of speech rhythm (e.g., Grabe & Low 2002), then you might consider using an automatic method of the kind mentioned earlier. Even if the boundaries have not all been accurately placed using the automatic procedure, it is still generally quicker to edit them subsequently rather than to place boundaries using manual labeling from scratch. As far as manual labeling is concerned, it is once again important to adhere to guidelines, especially if the task is carried out by multiple transcribers. There are few existing manuals that provide any detailed information about how to segment and label to a level of detail greater than a broad, phonemic segmentation (but see Keating et al. 1994b for some helpful criteria in providing narrow levels of segmentation and labeling in English spontaneous speech; and also Barry & Fourcin 1992 for further details on different levels of labeling between the acoustic waveform and a broad phonemic transcription). For prosodic annotation, extensive guidelines have been developed for American and other varieties of English as well as for many other languages using the tones and break indices labeling system: see for instance Beckman et al. (2005) and other references in Jun (2005).

Labeling physiological data brings a whole new set of issues beyond those that are encountered in acoustic analysis because of the very different nature of the signal. As discussed in Chapter 5, data from electromagnetic articulometry can often

be annotated automatically for peaks and troughs in the movement and velocity signals, although these landmarks are certainly not always reliably present, especially in more spontaneous styles of speaking. Electropalatographic data could be annotated at EPG landmarks such as points of maximum tongue–palate contact, but this is especially time-consuming given that the transcriber has to monitor several contacts of several palatograms at once. A better solution might be to carry out a coarse acoustic phonetic segmentation manually or automatically that includes the region where the point of interest in the EPG signal is likely to be, and then to find landmarks like the maximum or minimum points of contact automatically (as described in Chapter 7), using the acoustic boundaries as reference points.

Once the data has been annotated, then it is important to carry out some form of validation, at least of a small, but representative, part of the database. As Schiel and Draxler (2004) have noted, there is no standard way of doing this, but they recommend using an automatic procedure for calculating the extent to which segment boundaries overlap (they also point out that the boundary times and annotations should be validated separately although the two are not independent, given that, if a segment is missing in one transcriber's data, then the times of the segment boundaries will be distorted). For phoneme-size boundaries, they report that phoneme boundaries from separate transcribers are aligned within 20 ms of each other in 95 percent of read speech and 85 percent of spontaneous speech. Reliability for prosodic annotations is somewhat lower (see e.g., Jun et al. 2000; Pitrelli et al. 1994; Syrdal & McGory 2000; Yoon et al. 2004 for studies of the consistency of labeling according to the tones and break indices system). Examples of assessing phoneme labeling consistency and transcriber accuracy are given in Pitt et al. (2005), Shriberg and Lof (1991), and Wesenick and Kipp (1996).

1.3.7 Some conventions for naming files

There are various points to consider as far as file naming in the development of a speech corpus is concerned. Each separate utterance of a speech corpus usually has its own base-name with different extensions being used for the different kinds of signal and annotation information (this is discussed in further detail in Chapter 2). A content-based coding is often used in which attributes such as the language, the varieties, the speaker, and the speaking style are coded in the base-name (so **`EngRPabcF.wav`** might be used for English, RP, speaker **`abc`** who used a fast speaking style for example). The purpose of content-based file naming is that it provides one of the mechanisms for extracting the corresponding information from the corpus. On the other hand, there is a limit to the amount of information that can be coded in this way, and the alternative is to store it as part of the annotations at different annotation tiers (see Chapter 4) rather than in the base-name itself. A related problem with content-based file names discussed in Schiel and Draxler (2004) is that there may be platform- or medium-dependent length restrictions on file names (such as in ISO 9960 CDs).

The extension **`.wav`** is typically used for the audio data (speech pressure waveform) but other than this there are no conventions across systems for what the extensions denote, although some extensions are likely to be specific to different systems (e.g., **`.TextGrid`** is for annotation data in Praat; **`.hlb`** for storing hierarchical label files in Emu).

Schiel and Draxler (2004) recommend storing the signal and annotation data separately, principally because the annotations are much more likely to be changed than the signal data. For the same reason, it is sometimes advantageous to store separately the original acoustic or articulatory sampled speech data files obtained during the recording from other signal files (containing information such as formants or spectral information) that are subsequently derived from these.

1.4 Summary and Structure of the Book

The discussion in this chapter has covered a few of the main issues that need to be considered in designing a speech corpus. The rest of this book is about how speech corpora can be used in experimental phonetics. The material in Chapters 2 to 4 provides the link between the general criteria reviewed in this chapter and the techniques for phonetic analysis of Chapters 5 to 9.

As far as Chapters 2 to 4 are concerned, the assumption is that you may have some digitized speech data that might have been labeled and the principal objective is to get it into a form for subsequent analysis. The main topics that are covered here include some routines in digital signal processing for producing derived signals such as fundamental frequency and formant frequency data (Chapter 3) and structuring annotations in such a way that they can be queried, allowing the annotations and signal data to be read into R (Chapter 4). These tasks in Chapters 3 and 4 are carried out using the Emu system: the main aim of Chapter 2 is to show how Emu is connected both with R and with Praat (Boersma & Weenink 2005) and WaveSurfer (Sjölander 2002). Emu is used in Chapters 2 to 4 because it includes both an extensive range of signal processing facilities and a query language that allows quite complex searches to be made of multi-tiered annotated data. There are certainly other systems that can query complex annotation types of which the NITE-XML[18] system (Carletta et al. 2005) is a very good example (it too makes use of a template file for defining a database's attributes in a way similar to Emu). Other tools that are especially useful for annotating either multimedia data or dialogues are ELAN[19] (EUDICO Linguistic Annotator), developed at the Max Planck Institute for Psycholinguistics in Nijmegen, and Transcriber,[20] based on the annotation graph toolkit (Bird & Liberman 2001; see also Barras et al. 2001).[21] However, although querying complex annotation structures and representing long dialogues and multimedia data can no doubt be more easily accomplished in some of these systems than they can in Emu, none of these at the time of writing includes routines for signal processing, the possibility of handling EMA and EPG data, or the transparent interface to R that is needed for accomplishing the various tasks in the later part of this book.

Chapters 5 to 9 are concerned with analyzing phonetic data in the R programming environment: two of these (Chapters 5 and 7) are on physiological techniques; the rest make use of acoustic data. The analysis in Chapter 5 of movement data is simultaneously intended as an introduction to the R programming language. The reason for using R is partly that it is free and platform-independent, but also because of the ease with which signal data can be analyzed in relation to symbolic data, which is often just what is needed in analyzing speech phonetically. Another

is that, as a recent article by Vance (2009) in *The New York Times* made clear,[22] R is now one of the main data-mining tools used in very many different fields. The same article quotes a scientist from Google who comments that "R is really important to the point that it's hard to overvalue it." As Vance (2009) correctly notes, one of the reasons why R has become so popular is because statisticians, engineers, and scientists without computer-programming skills find it relatively easy to use. Because of this, and because so many scientists from different disciplinary backgrounds contribute their own libraries to the R website, the number of functions and techniques in R for data analysis and mining continues to grow. As a result, most of the quantitative, graphical, and statistical functions that are needed for speech analysis are likely to be found in one or more of the libraries available on the R website. In addition, and as already mentioned in the preface and earlier part of this chapter, there are now books specifically concerned with the statistical analysis of speech and language data in R (Baayen, in press; Johnson 2008) and much of the cutting-edge development in statistics is now being done in the R programming environment.

Notes

[1] www.ldc.upenn.edu/
[2] www.elra.info/
[3] www.elda.org/
[4] www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html
[5] www.cslu.ogi.edu/corpora/corpCurrent.html
[6] http://talkbank.org/
[7] www.mpi.nl/DOBES
[8] http://lands.let.kun.nl/cgn/ehome.htm
[9] www.ipds.uni-kiel.de/forschung/kielcorpus.en.html
[10] www.hcrc.ed.ac.uk/maptask/
[11] http://andosl.anu.edu.au/andosl/
[12] http://vic.psy.ohio-state.edu/
[13] www.cstr.ed.ac.uk/research/projects/artic/mocha.html
[14] www.maccs.mq.edu.au/~nwdb
[15] Websites that provide helpful recording guidelines are those at Talkbank and at the Phonetics Laboratory, University of Pennsylvania: www.talkbank.org/da/record.html, www.talkbank.org/da/audiodig.html, and www.ling.upenn.edu/phonetics/archive/FieldRecAdvice.htm.
[16] Florian Schiel, personal communication.
[17] See www.phonetik.uni-muenchen.de/Bas/software/speechrecorder/ to download SpeechRecorder.
[18] http://sourceforge.net/projects/nite/
[19] www.lat-mpi.eu/tools/elan/
[20] http://trans.sourceforge.net/en/presentation.php
[21] Plans are currently in progress to build an interface between ELAN and Emu annotations. There was an interface between Transcriber and Emu in earlier versions of both systems (Barras et al. 2001; Cassidy & Harrington 2001). Since, at the time of writing, Transcriber is being redeveloped, the possibility of interfacing the two will need to be reconsidered.
[22] My thanks to Andrea Sims and Mary Beckman for pointing this out to me. The same article in *The New York Times* also makes a reference to Emu.