Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases

Helen Fraser

School of Languages Cultures and Linguistics, University of New England

ABSTRACT This article considers the reliability of transcripts used as evidence in court, especially transcripts of poor recordings. Background information about human speech and speech perception is presented, and the implications of this information for the use of transcripts of different kinds in legal contexts is considered. Finally, recommendations are made to allow judgement of the reliability of existing transcripts, ensure that newly created transcripts are reliable, and to ensure that transcripts are presented to a jury appropriately.

KEYWORDS transcription, forensic phonetics, human speech perception, transcript reliability

INTRODUCTION AND OVERVIEW

Transcripts are used in court in two main ways: as a record of proceedings, and as evidence in cases. This article concentrates on the latter, though the former are also discussed briefly. Transcripts used as evidence come from a variety of sources, commonly including intercepted telephone calls or undercover recordings made by listening devices. Disputes as to whether a transcription accurately represents what was said when the recording was made arise most often in cases where use of a listening device has resulted in a very poor quality recording. This article is motivated by consideration of the number of cases in which such disputes have been conducted with insufficient awareness of linguistic, phonetic, and psycholinguistic issues, and without recourse to advice from specialists with relevant expertise in linguistics.

This is a situation not unfamiliar to linguists, simply because the discipline of linguistics is not as widely recognized and understood as other areas of expertise. If an investigating team find an unknown substance at the scene of a crime, they do not give it to a passing lay person for analysis, or analyse it themselves, even if they are quite sure what it is. They send it to a chemist or other appropriate expert, since the court expects to hear a scientific analysis from an independent expert. Faced with a barely audible but possibly incriminating recording, however, all too often, the police transcribe it themselves, perhaps sending portions of it for verification to a sound engineer – an expert, but, as I will argue below, the wrong kind of expert for this particular task.

1350-1771
Forensic Linguistics 10(2) 2003

This article is therefore intended not as a presentation of new research findings, but as a tutorial statement about issues affecting the reliability of transcripts for use as evidence in court. Since transcription depends upon speech perception, the article first reviews some important aspects of human speech and speech perception. It then draws out the implications for the reliability of transcripts, and finally sets out several recommendations for making and using transcripts so as to maximize their reliability. The overall conclusion of the argument is that poor quality recordings can often be transcribed with considerable reliability, but only if certain strict guidelines are followed in their making, including guidelines for determining which kinds of material are not transcribable to a level suitable for use in court. The article also points out that the creation of a good transcript is only a first step in the use of recorded speech as evidence. Even with a reliable transcription, substantial issues can arise around the interpretation of the speakers' meaning and intention. Weighing these issues also requires the use of linguistic expertise.

ASPECTS OF HUMAN SPEECH PERCEPTION

In this section I introduce two aspects of human speech perception which, while very well established in the research community (Fillenbaum 1971, Osgood and Sebeok 1965, Saporta 1961, Berko-Gleason and Bernstein Ratner 1993), are little known to 'ordinary people'. These have to do with the contribution to perception made by, first, the perceiver, and second, the nature of speech itself.

The unacknowledged role of the perceiver

'When we listen to someone speaking, what we hear depends on the sounds that the person utters.' This sounds so obvious as to be a truism – but actually it is a very partial truth. Accepting it on face value obscures, rather than reveals, some important and relevant facts about speech perception.

In fact, when we listen to someone speaking, what we hear depends on three things:

a) the sounds that they utter;
b) the context of other sounds in which the particular sounds to which we are attending are uttered; and
c) the listener's knowledge and expectations about the language the speaker is using, and the situation in which they are speaking.

Strangely, however, though perception is heavily influenced by all three of these factors, listeners themselves are highly aware of the role of (a), the sounds themselves, and generally do not notice at all the role of (c), their own perceptual activity. People have the overwhelming impression that the speaker's sounds simply contain the message, unequivocally. They

think that speech is 'clear' if the sounds are clear, and 'unclear' if they are not; whereas in fact perceptual clarity depends upon an interaction between all three of the factors above.

Although we generally don't notice our own contribution to perception, speech perception is an active, rather than a passive, process, with the hearer actively constructing, rather than passively picking up, the speaker's message. Certain kinds of experience demonstrate the listener's constructive activity, and give useful information about the nature of that activity.

Consider first the experience of listening to someone speaking a foreign language. If you have no knowledge of the language, it is not just that you can't understand the words – you can't even hear the words properly, or repeat them accurately. You can't parse the signal (break it into words and phrases), or remember it for more than a few seconds. To someone who does know the language, on the other hand, the signal is perfectly clear and intelligible. Even if there are some words they don't understand, they at least know where those words begin and end, and can remember them well enough to ask 'What does X mean?'.

The sound – or acoustic signal – has remained the same in both cases; what has changed is the listener, specifically the listener's knowledge of the language.

Even in listening to a well-known language, the listener's knowledge, in the sense of ability to predict what is likely to be said, has a strong effect on how 'clear' the speech sounds. Thus you can be listening to someone speak and finding them perfectly clear and intelligible – until they say something that is not predictable. Consider the following real life example.

Two colleagues start chatting as they walk towards the coffee-room. As the first one enters the room he says 'I see the sink's back then?'. Whereas everything else he had said up till then was perfectly clear and unambiguous, his companion found this sentence quite unintelligible, even after several repetitions. The reason was that the speaker knew that there had been a plumbing problem recently which meant that the sink in the coffee room had had to be removed. The listener did not know about this, and since the sink was now back as before, could not predict, or even realize with hindsight, that there would be any issues about the sink to discuss.

The point is, the colleague did not say, 'What do you mean, the sink's back?'. It wasn't just that he did not understand the *meaning* of the sentence; he really couldn't hear the sounds properly. Even repeating the sentence more clearly did not help. It was only when the speaker realized the listener didn't know about the sink and explained the background, that the listener understood the message as a whole, and the individual words within it.

The relationship between predictability and clarity is seen especially

strongly when there is significant background noise: noise affects intelligibility far more if the speech is unpredictable, and far less if the speech is highly predictable. Classic psycholinguistic experiments investigated this relationship by playing subjects recordings of speech masked by noise (Miller 1954). One variable was how much noise was masking the speech; the other was what kind of message the speech was conveying: a series of random words; a sensible, commonplace sentence; or a nonsensical sentence. A sensible, commonplace sentence remains intelligible at much higher levels of background noise than the other conditions – precisely because the hearer can use the predictability of its grammar and meaning to help in interpreting it.

An everyday experience that shows the same thing is that of talking on the telephone. Telephone transmission cuts out a great deal of acoustic information that in principle is required to differentiate the sounds of English. For example, a straightforward test of the discriminability of individual sounds shows that sets of sounds like 's, f, th', or like 'm, n, ng', are hard to distinguish over the telephone (due to the loss of certain frequency bands).

However, when we have a conversation on the telephone, we understand each other perfectly well, not even noticing that these sounds are difficult to discriminate. All is well until – but only until – the level of predictability of the speech decreases, for example when we mention names, which are generally less predictable than other words. This, of course, is the reason that we have to spell out people's names on the telephone, and why when we do, we use devices like 'm for mother, f for freddy, s for sam': only when a sound is embedded in a meaningful word, is it predictable enough to be clear (i.e. the listener knows it couldn't be 's for sreddy' or 'f for fam'). Note that for this to work, we have to choose our example words carefully: saying 's for sing' might not work – because of the equally common word 'thing' we would have to say 's for sing a song'.

Examining experiences like these helps us to realize the active role we play in constructing the messages we hear by combining the information in the speech signal with the knowledge in our heads.

The misunderstood role of the acoustic signal

We have emphasized the role of the listener, but we must not thereby negate the role of the signal. The listener is not free to hear any old message just because it is what they expect to hear, regardless of the sounds that are actually uttered. If someone says 'roses are red', it is not open to the listener to hear 'a funny thing happened to me on the way to the lecture theatre' – even if the latter was what the listener predicted the speaker would say. It is very important to stress that the sounds that are uttered certainly do matter – they just don't play quite the role in perception that non-linguists usually think.

The listener's interpretation does rely heavily on the sounds in the acoustic signal, but the acoustic signal does not fully determine the listener's interpretation. Rather, the information in the acoustic signal provides one of several sets of clues, which the listener uses in conjunction with other, non-acoustic, clues, to figure out the message. It is common to use the terms 'top-down information' to refer to the clues that come from the listener's knowledge and prediction, and 'bottom-up information' for the clues that come from the signal itself.

In considering the role of the acoustic signal itself, it is necessary first to overcome some common misunderstandings about the nature of speech. Most people without special training in phonetics make an unconscious assumption that speech is a little like the print on this page: that it consists of a small number of sounds, repeated in varying combinations, so that each sound is distinct from its neighbours, and each word is separated by a small but easily discernible space. In fact this is a prejudice that comes to us through the learning of alphabetic literacy (Olson, Torrance and Hildyard 1985). The real nature of speech is very different, as we will see.

Of course, most people are aware that speech is not like writing in the sense that English spelling is highly irregular, with an unsystematic letter-to-sound relationship. Thus there is no letter to represent the sound at the beginning of 'then' as opposed to the sound at the beginning of 'thin'; 'cough' has five letters but only three sounds, and a pronunciation quite different from 'bough' despite their similar spelling.

With a small amount of training, people can learn to do 'phonemic transcription', which systematically uses one special symbol for each distinctive unit of sound, or phoneme. A phoneme is an individual sound of a particular language. For instance, the 's' of 'sat', or the 'f' of 'fat', the 'i' of 'sit', the 'th' of 'then' and the 'th' of 'thin' are all phonemes of English. Each language has its own set of phonemes (or equivalent units – some languages have a structure that makes them difficult to describe strictly in terms of phonemes). Some alphabetic writing systems (such as that of Spanish) represent the phonemes of their languages reasonably unambiguously, as opposed to English, in which the spelling is irregular – though not nearly as chaotic as is sometimes thought (Carney 1997, Venezky 1970).

It is important to realize, however, that though phonemic transcription does overcome some of the irregularity of English spelling, it does not thereby give a true representation of the real nature of speech. A phonemic transcription is merely a modified orthography, and still represents speech using a small set of discrete units arranged in varying combinations, with small gaps between the symbols and larger gaps between the words. This is not at all what real speech is like; as we will see, it is continuous. Thus there is a fundamental distinction between a phonemic, or, more generally, a phonological, transcription, which represents speech in terms of its distinctive units, and a phonetic transcription, which aims to represent

speech in terms of its actual sound (e.g. Clark and Yallop 1990, Ladefoged 1993, Spencer 1996).

It is important to stress this distinction here because some who have studied a little linguistics, enough to have learned special symbols for the phonemes of English, believe that a phonemic transcription gives a true representation of speech, even calling a phonemic transcription 'phonetics' (an easy confusion, since the phoneme symbols are drawn from a set called the International *Phonetic* Alphabet, or IPA).

Phonemic transcription is the very first step in learning phonetics. It is an important first step, and a useful skill for many purposes. However it is very much just that – a first step. In order to understand speech perception, and the nature of transcription, and especially the constraints on the accuracy of transcription, it is essential to go very far beyond a phonemic understanding of speech.

THE NATURE OF SPEECH

Importance of stress and rhythm

Psycholinguists and phoneticians have investigated the characteristics of the speech signal which are important to perception through a wide range of classic experiments. For example, through experimenting with the perception of speech in noise as described above, a good deal is known about which aspects of speech are 'robust' (easy to distinguish even with heavy background noise), and which degrade significantly even with minimal background noise; and also about which aspects of speech are liable to become confused with which other aspects when background noise is significant (Miller and Nicely 1955). It is from these experiments that we know, for example, that voiceless fricatives like 's' and 'f', and nasals like 'm' and 'n' are highly confusable, whereas stressed vowels are fairly robust. We also know that the overall rhythm of speech is extremely significant in speech perception, much more so than most individual sounds. The individual sounds are of course important as clues, but far less so than the overall rhythm – they can be over-ridden by the perceptual mechanism (again within constraints), especially if other clues seem to suggest a conflicting interpretation.

One area of study which has been very useful in investigating these issues is research on hearing errors. These are the hearing counterpart of the well-known phenomenon of 'tips of the slung', or slips of the tongue (Fromkin 1973, 1980). These 'speech errors' occur when a speaker makes transient errors of performance, for example when the weather forecaster says 'widespread fosts and frogs' instead of 'widespread fogs and frosts'.

Hearing errors, by comparison, are mistakes made by the hearer rather than the speaker, and, like speech errors, are often the subject of jokes, (e.g. the old chestnut about the soldier who relays the commandant's

message as 'send three and fourpence we're going to a dance' instead of 'send reinforcements we're going to advance'). As well as being amusing, hearing errors can give a great deal of information about the process of speech perception and the factors that affect it (Bond 1999).

One of the most important lessons from the study of hearing errors is the enormous influence on speech perception of rhythm and stress, as opposed to individual sounds. The above example is a perfect illustration of this: while the stressed syllables in the two versions are very similar ('send', 'for/four', 'dance/advance'), the unstressed syllables are not just quite different in sound, but crucially different in meaning and grammar. For example, 'reinforcements' is one word, 'three and fourpence' is three; 'advance' is a verb, 'a dance' is a noun, and so on.

Consider the list of real-life hearing errors in Table 1, paying attention to the relationship in sound between what is said and what is heard.

Table 1 Examples of hearing errors

What was said	**What was heard**
I'm a student too, I'm not just a wife	I'm a student too, in Manchester
Got a notebook handy?	got an opal candy?
maple leaf	make believe
but lizards don't even have teeth	at least when it's finished we can have tea
I think I see a place	I think I see his face
this report is tolerable	this report is horrible
Australians all let us rejoice	Australia's only ostriches
the girl with kaleidoscope eyes	the girl with colitis goes by
all staff email	all star female
gladly thy cross I'd bear	gladly the cross-eyed bear
this guy's in love with you	the sky's in love with you
when the going gets tough	go and get stuffed

The last few cases in Table 1 are examples of a particular class of errors called 'mondegreens' – plausible but incorrect parsings, often but not always by children, of slightly arcane phrases (often from songs). Whereas hearing errors are transient, mondegreens can last for years, even a lifetime – many people can tell of misinterpretations they only discovered to be errors well into adulthood. An internet search on 'mondegreen' brings up a wealth of examples and information compiled by enthusiasts for this type of language play.

What we see in both hearing errors and mondegreens is further evidence of the role the hearer plays in constructing a plausible message using clues in the speech signal, rather than simply picking up the sounds and translating them into a message.

Further, we gain information about the types of clues in the speech signal that are most important to perception. In each case, the similarities between what is said and what is heard are in the stressed vowels and in the overall pattern of rhythm and intonation. The details of individual sounds, especially in unstressed syllables, can be radically different between what is said and what is heard, though the patterns of confusability are constrained in ways that have been studied in some detail (e.g. Bond 1999).

Continuous not discrete

It seems, then, that when we listen to someone talk, we pay most attention not to the individual sounds, but to the overall pattern of their speech.

This is true not just of 'deviant' or error-full perception, but also of successful communication. Another type of experiment that psycholinguists have used extensively to investigate speech perception is called 'gating' (Shockey 2003). In gating, a sentence or conversation is recorded, and then played back in very short sections, and the effect on subjects' perception noted.

The most dramatic observation is that the individual sounds of speech are quite unintelligible when heard on their own. Even whole words and phrases are radically misinterpreted, or cannot be interpreted at all, when excised from longer stretches of speech (Pollack and Pickett 1963). The point is that, until we have some context to constrain our predictions and expectations, we cannot interpret individual sounds. The ear needs a sufficient duration of speech to give a context for interpretation.

The effect of a gating demonstration is difficult to convey on paper (but see Fraser 2001 which includes several audio demonstrations). For the sake of written presentation, a close analogy can be given from handwriting. Small sections cut out of a perfectly intelligible written text can be quite indecipherable. Consider the examples in Figure 1.

Figure 1 Several sections excised from a continuous handwritten message

Most people find it difficult to be sure even of how many letters are represented by such scrawls, let alone what the letters are. There is insufficient context to let us use our top-down knowledge to aid our interpretation. We see, then, that only some of the clues we need to construct a message are contained in each individual letter.

Consider the example in Figure 1 again: Does it help to know that the sections were excised from an address on an envelope? That the first scrawl came at the end of a line in the address? When you are ready, turn to the end of this article to see the excised sections in their original context.

Now look again at the scrawls above – they have probably become much clearer now that you know the context from which they came, and can apply top-down information to their interpretation. This is exactly what happens in a gating experiment in speech: once we have heard the full context, when we go back to the excised individual sounds – which could not be interpreted at all the first time they were heard – they have become quite clear. In fact it becomes amusing to see others struggling with their interpretation.

It may be tempting to complain that the handwriting in this example is particularly messy. Note though that the original was comprehensible in its context on an envelope (the writing may not be good, but the letter was delivered!). The messiness of the individual letters may make the writing difficult for a child or a foreigner to understand, but to someone who is a mature reader, and knows the language, conventions of addresses, and the names of local streets and suburbs, the handwritten address *as a whole* is far more comprehensible than the excerpts. Such a person uses the contextual information to guide their perception of the 'bottom-up' data.

Two things are worth pointing out. First, we have already said that speech is not like printed language. Handwriting presents a better analogy, though even clear speech is 'messier' than messy handwriting. At least in handwriting we generally maintain a gap between one word and the next, whereas in speech even the words run into one another.

Second, the recognizability of individual letters in print, as opposed to handwriting, is something of a red herring. Even print, which does consist of a sequence of discrete symbols separated by spaces, is not read letter-by-letter. Although we are taught to read by assigning a sound to each letter and 'sounding out' words, this is not how fluent reading works. It is true that any one letter excised from the text you are currently reading is individually recognizable, but it is certainly not true that in reading this text you are individually recognizing each letter en route to understanding the words. Although the 'sounding out' stage is essential to learning literacy, to continue to read by recognizing each individual letter is to have a serious reading disability (Byrne 1998, Just and Carpenter 1987).

To be able to read fluently we must learn to recognize global patterns of words and phrases, and to use our knowledge and understanding to predict what is coming next. The same is true of speech.

Interestingly though, once we have learned the basics of reading, we usually get so involved in the topics we are reading about that we do not stop to question our teachers' assertions that letters represent sounds, and words are logical sequences of letters representing logical sequences of sounds. Thus a by-product of being taught to read is that we grow up believing that speech really is, or should be, a sequence of discrete sounds such as those on a printed page. Un-learning this 'obvious fact' is a major task for undergraduate linguistics students.

Non-invariance

We have seen that people generally discount their own contribution to the construction of the messages in speech, and think all the information is actually in the speech signal itself; and also that most people have quite strong beliefs about what speech itself is like, heavily influenced by their knowledge of written language. As well as thinking that each word is separated by a short pause and clearly distinct from its neighbours, and that each sound is discrete like a letter, another belief many people hold is that individual sounds (or phonemes) are *invariant*. Printed letters are invariant (at least within a particular font), in the sense that every occurrence of the letter 'b' (for example) is the same as every other occurrence of 'b', and that every time the letter shape 'b' is encountered, we can be sure it represents the letter 'b'.

We have seen that spoken language is much more like handwriting – and like particularly scrawling handwriting at that. In even the most careful and correct speech, individual sounds merge into one another, adapting themselves to the preceding and following sounds. In the process they lose not just their discreteness, but also their invariance. So, just as two instances of the same letter can look quite different (compare the two 't's in the *Kennedy St Kingston* example in Figure 2), so two instances of the same sound can be acoustically quite different. Not only that, but, just as one scrawl can be interpreted as one letter in one context and another in another (see how the 'o' of *Kingston* could be an 'e' in another context), so particular sections of acoustic signal can be interpreted as quite different sounds depending on the context.

The result is that, despite our strong impression that all the 'b' sounds we hear are the same, acoustically there are many forms of 'b' – and of every other sound. In fact, the examples from handwriting we have looked at are rather superficial in relation to the invariance of sounds of real speech examples. It is quite possible to confidently hear a phoneme in speech which acoustic analysis shows simply not to be there – or vice versa (for example, most people believe that the word 'prints' contains a 't'

sound, whereas 'prince' does not, but in fact these two words are very often pronounced identically, from an acoustic point of view).

The main point, however, in the present context, of these handwriting examples, is that accurate perception of one or two letters depends on having enough context for top-down information to be used in its identification. A dispute over the identity of an individual letter cannot be settled purely by analysis of that letter alone, no matter how extensive that analysis is, or how great the technical expertise of the analyst (would it be possible to prove that the 'o' of *Kingston* is an 'o' rather than an 'e' without looking at the word and the phrase as a whole?). The same is true of speech. Before considering the implications of this for the use of transcripts as evidence in court, however, it is worth pursuing the general background on speech just a little further.

SOURCES OF ERROR IN PERCEPTION

How speech perception works

A good way of thinking about speech perception is as a search for clues in the acoustic signal, with these clues then being used, along with other information, to construct the message, or deduce what it must have been. The clues from the signal (bottom-up) are located not only in the individual sounds, but in the overall rhythm of speech. As we have seen, speech is not perceived sound-by-sound, but requires integration over a sufficient time period to enable the pattern of stress and rhythm to be apprehended properly. The top-down clues are of two main kinds. The first is information the listener brings about the context in which the speech occurs (e.g. what it is about). The second is information contained in the language itself (e.g. whether the next word is likely to be a verb or a noun). Both these types of top-down information aid the listener by helping predict what is likely to be said.

One implication of this view is that perception of speech can be manipulated. Since the perceiver does a job of deduction based on clues, perception can be 'tricked', either deliberately or by chance.

Inducing errors by manipulating the signal

One way of manipulating a perceiver's interpretation of sounds is to change the acoustic context in which they occur – ie. to change, not the sounds in focus, but the sounds nearby. An elementary example which never fails to amaze those who see it for the first time is to record a word like 'spat', and then offer to play it without its 's'. Most people predict unhesitatingly that the result will sound like 'pat' – but in fact it sounds like 'bat' (see Fraser 2001 for a demonstration). It is important to emphasize the strength of this perception. With the 's' attached, the sound is absolutely clear as the 'p' of 'spat'; without the 's', the sound is equally

unambiguous as the 'b' of 'bat'. The normal pronunciation of 'p' in words like 'spat' (i.e. after an 's') is in fact identical to the normal pronunciation of 'b' in words like 'bat'. Yet because they are represented with different letters, we simply assume they must be different sounds.

Another demonstration involves splicing the same acoustic signal into two or more different acoustic contexts, again with remarkable effects on perception. Take the old jingle, 'I scream, you scream, we all scream for ice-cream'. The joke of the jingle is created by the acoustic similarity of 'I scream' and 'ice-cream' – but what is often not realized is that the two phrases are more than just similar. Now it is true that it is perfectly *possible* to say 'I scream' and 'ice-cream' so that they sound quite different. It is also true that (in many dialects, including Australian English) the two phrases are *often* pronounced, even in clear, well-educated speech, so as to be functionally indistinguishable – despite small differences of acoustic detail.

In these dialects, it is possible to record *one* of these phrases (either 'ice-cream' or 'I scream') and splice it *twice* into either side of the word 'for' – so that the *exact same acoustic material* occurs on either side of the word 'for'. Playing the result gives the very strong percept of 'I scream for ice-cream' and doesn't sound at all like 'ice-cream for I scream', or 'Ice-cream for ice-cream', or 'I scream for I scream'. If it is played to someone who hasn't actually seen this done in front of their eyes, that person will find it difficult to accept that the two phrases are identical, and will point out 'very subtle' (in fact, imaginary) differences between them. The point is that in this case these subtle differences must be created by top-down, rather than bottom-up, perceptual processes – a nice demonstration of not just the power of top-down processing, but of speakers' unawareness of them.

These are not just isolated examples. The same thing can be done with many other phrases, for example 'This guy' can be spliced before 'was a beautiful shade of blue' to create the percept 'The sky', or 'a few' can be spliced into 'I'll have a few ___ have a few' to create the percept 'if you'. The point, again, is that in speech, all is not as it seems to be. In order to understand it well, many 'obvious facts' have to be re-evaluated, and many counter-intuitive truths to be accepted. Considerable training in linguistic phonetics is needed to achieve this re-evaluation, and to understand its implications.

Inducing errors by manipulating the perceiver

Another way of manipulating speech perception is by affecting the listener's knowledge or predictions about the message. This has been investigated experimentally, again in classic studies of psycholinguistics. We have already mentioned the experimental technique of playing recordings of sentences masked by noise. One early application of this

method (Bruce 1958) showed that subjects' perception of such sentences can be greatly influenced by telling them something about the content of the sentences – enabling them to use top-down information.

Subjects heard the same set of recorded sentences, masked by the same noise, several times. The sentences were like these:

Sentence 1. I tell you that our team will win the cup next year.
Sentence 2. You said it would rain but the sun has come out now.

On each hearing, each sentence was preceded by a word which the subjects were told was the topic of the sentence (e.g. 'sport', or 'weather'). Subjects had to repeat back what they thought the sentence was, and give a rating of their confidence.

The original intention of the experiment was to show that having a hint as to the general topic of the sentence aided perception significantly. This was indeed demonstrated: subjects who heard 'sport' before Sentence 1 interpreted it much more successfully than those who heard 'weather' before the same sentence; those who heard 'weather' before Sentence 2 interpreted it much more successfully than those who heard 'sport', and so on. This shows the degree to which contextual information beyond the mere acoustic signal is involved in perception.

What was also discovered, however, to the surprise of the experimenter at the time, was that subjects very often interpreted a sentence in line with the topic word, even if that was not the right topic for the sentence. For example, those who heard 'food' before Sentence 1 (which was actually about sport) reported hearing sentences about food. Thus, one subject reported hearing Sentence 1 as:

Sentence 1 (food): I tell you that I feel more hungry than you are.

Those who heard 'travel' before exactly the same sentence (about sport), masked by exactly the same degree of noise, reported hearing sentences about travel, for example:

Sentence 1 (travel): I tell you that I too will leave next year.

The design of the experiment allowed equally unexpected, but even more significant, observations to be documented. In fact, each subject heard the same set of five sentences repeated five times, with only the topic-words changing. Yet they thought they had heard five different sets of sentences. Here is an example of one subject's interpretation of the 'sport' sentence (1) under different topic-word conditions:

Sentence 1 (weather): I tell you that I see the wind in the south next year.
Sentence 1 (health): I tell you that our team has been free from injury all this year.
Sentence 1 (food): I tell you that our tea will be something to do with beer.

Most importantly, each subject was generally quite happy with his or her interpretation. They were unaware that any other interpretation was possible – not realizing that other participants, with other topic-information, were equally happy with their own, quite different, interpretations, and not even realizing that they themselves had interpreted the same sentence quite differently when they had heard it with a different topic-word.

Error and accuracy

We have talked a good deal about the errors to which perception, and the analysis of speech, are prone. It is most important now to emphasize that all this does not mean that perception is necessarily unreliable. The errors we have mentioned are *possible* errors in the *spontaneous perception* of speech, which become particularly evident when we stress the perceptual system in order to study its error patterns; they are not *inevitable* errors in *considered interpretation* of speech.

After all, though speech perception is prone to all the errors just discussed, especially under non-optimal conditions, we do in fact generally succeed in communicating with one another. Even when we talk to each other in noisy environments, we can usually understand quite well, due to a phenomenon known as the 'cocktail party effect', which enables the perceptual mechanism to focus on one line of speech and ignore all others.

The reason for our normally effective perception is that in face-to-face communication we know how to judge the accuracy of our perception, how to question it if it is doubtful, and how to correct it if it is inaccurate.

These are exactly the steps that are necessary in creating accurate transcripts. The problem is that in transcribing from a recording we are not in an ordinary communicative situation, with a meaningful context, and the speaker present to correct any important errors. Rather we are abstracted from the real situation, with greatly reduced top-down information, and, in the case of a poor recording, reduced bottom-up information too. We have seen that in such situations, the correlation between confidence and accuracy can be very low. That is why specialized knowledge is needed to understand the factors that affect the reliability of perception.

JUDGING THE ACCURACY OF TRANSCRIPTS

It is clear that the first step in transcribing recorded speech is the perception of the linguistic structure of what is said. Thus the factors that affect perception also affect transcription.

It is important to emphasize, however, that nothing we have said about perception need suggest that transcription is necessarily unreliable. It is quite possible to make a reliable transcription. The problem is it is also possible to make an unreliable transcription. What is needed is a reliable statement of the degree of confidence that should be placed in any particular transcript. For this, understanding the fact that perception is prone to error, knowing about common types of error, and being aware of the factors that can affect the achievement of accurate transcription is essential. This understanding is only available to someone with considerable expertise in phonetics and psycholinguistics.

In this section we consider some relevant issues, distinguishing two different types of transcription, first transcription of relatively good recordings of relatively clear speech, and second, poor recordings of unclear speech. As anyone who has ever tried transcribing speech from a recording will know, even when the speech is quite clear, transcribing it is an extremely painstaking process, requiring a time expenditure many times the duration of the speech itself.

Transcripts of clear recordings

Normally, transcriptions used for important purposes (e.g. courtroom or parliamentary proceedings, or intercepted telephone conversations) are of portions of speech of substantial duration, recorded on appropriate equipment in good conditions. In such cases, a sufficiently accurate transcription can be made by any literate – and patient – person.

Even the best such transcript however, will only be *sufficiently* accurate, not a hundred per cent accurate. This is for two main reasons. The first has to do with the differences between spontaneous speech (ie. speech produced 'on the fly'), and speech read aloud from a text. Spontaneous speech is not neatly structured according to the rules of grammar – and this is true even for people who speak very correctly and grammatically, whether or not they like to admit it (Brown 1977, Haberland 1994). In spontaneous speech, people frequently change their minds half-way through a sentence; they frequently add little asides under their breath; they frequently convey part of their message through gesture or implication rather than making everything explicit in words; they use many contractions and colloquialisms. This is true, not just for casual conversational speech, but also for prepared or formal speech. In fact it is the reason we generally prefer to listen to spontaneous speech rather than a read text. All the 'imperfections', strangely enough, serve to make spontaneous speech easier to understand, not more difficult, than a read text – unless the text has been specially prepared by someone with the rare skill of writing for reading aloud.

When we transcribe spontaneous speech we convert it into a written text and – to a greater or lesser extent, but always to some extent – adapt

it to be readable as a text. For this reason, transcribers generally omit 'umms and ahhs', separate out overlaps and interruptions, correct 'slips of the tongue', and 'tidy up' the grammar and pronunciation (e.g. rendering 'wanna' as 'want to'), so as to more accurately reflect the intention of the speaker (see Eades 1996, Walker 1990). If they don't do this, the transcription is very difficult to read – as are, for example, the very detailed transcripts used in linguistic studies of conversation (Wray, Trott and Bloomer 1998).

This is interesting because it once again reflects something that is familiar to psycholinguists and phoneticians: the fact that in perceiving speech, we 'edit out' most of the speaker's unintentional errors. We talked above briefly about 'slips of the tongue'. When these are humorous, or otherwise affect meaning, they will be commented upon. But the vast majority of slips of the tongue are not even noticed, or if they are, they are simply corrected by the listener without comment. In fact, experiments have been done (Tent and Clark 1980) in which deliberate errors were introduced into recorded speech, for example substituting all the 'p's with 't's. When such speech is played to subjects, they notice very few of the errors, preferring to go with the overall flow of meaning rather than pay attention to every detail of pronunciation.

The second reason transcription is only ever sufficiently accurate has to do with the nature of speech itself. We saw above that the most important clues in English are contained in the stressed words and syllables. The unstressed parts of speech are by definition less important and speakers sensibly choose not to waste too much time agonizing over the careful enunciation of the unstressed parts, leaving it to the listener to use top-down contextual information to interpret them. This causes no problems in face-to-face communication, both because considerable contextual information is available, and also because the speaker is available to clarify things if the listener gets lost (recall the 'sink' example from above).

However a transcriber is in a different situation. The context is not so immediately evident, and the speaker is not available for clarification. Mostly, this doesn't make much difference to the overall meaning – for example 'if you like' is often transcribed as 'if you'd like', or vice versa. Sometimes, however the differences can be quite significant, especially if the transcriber does not share the same knowledge as the participants in the original conversation – for example in transcriptions of my own talks, the word 'phonological' has been transcribed as 'psychological'. Many similar examples can easily be found in transcripts of very good, clear recordings such as ABC radio broadcasts, in Hansard or in transcripts of courtroom proceedings. From the point of view of speech perception, this is highly unsurprising – it simply shows the influence of the listener's knowledge on the way speech is perceived.

Finally we should consider a slightly more technical reason that even transcription of clear speech is not one hundred per cent accurate. Phonetically, speech is a continuous stream of sound, which is only roughly represented by standard orthography or even by phonetic transmission. However only for very specialized purposes do we bother to make a phonetic transcript – for most 'verbatim' transcripts, the aim is to capture the gist of what was said, to a level that various parties can agree upon, rather than every phonetic detail of the speech.

It is for all these reasons that it is said by those who have experience with transcription that it is virtually impossible to get a hundred per cent accurate transcription of a recording of any length. Every time you listen to it, you hear something new, and get a little bit more accurate. At a certain point, the returns on repeated listening are so small as to be not worthwhile – where this point comes depends, of course, on the purpose of the transcription, and the level of agreement between the parties who have to sign off on it as an accurate record of what was said. For most purposes a reasonably accurate transcript is perfectly adequate, and few disputes are likely to arise, since for most purposes it doesn't really matter whether the speaker said, for example, 'I'm gonna' or 'I am going to'.

Transcripts of poor recordings

Let us turn now to the question of transcription of difficult material – recordings made in a very noisy environment, with poor equipment, or in poor recording conditions (e.g. with a great deal of background noise, the speakers a long way from the microphone, incorrect settings for recording, etc). These are the cases where significant disputes can arise as to what words were spoken, and where important legal consequences can depend upon the reliability of a transcript.

Obviously it will be harder to transcribe difficult material than a clear recording. The level of accuracy that can be obtained will depend on a number of factors, which can be divided into those to do with the recording, and those to do with the transcriber.

Factors relating to the recording

The first factor is the exact nature of the noise or interference. For example a consistent background noise can be progressively ignored by the perceptual mechanism, allowing the transcriber to focus on the speech itself. Note that in some cases a consistent hum, hiss or buzz can be filtered out mechanically; but if it happens to be in the middle of the speech frequency range, filtering it out mechanically risks distorting the acoustic structure of speech; the ears are usually the best filtering devices available for this type of work.

Other types of interference are more difficult to filter out, either mechanically or perceptually. One of the worst is other speech, additional

to the speech being transcribed. For example if the recording is made in a crowded room, or with a television or radio playing nearby, transcription will be very difficult. We have seen that in face-to-face conversation, the perceptual system allows us to focus on one line of attended speech even under very difficult circumstances, such as a cocktail party. A tape recorder unfortunately has no such powers; it records all noise equally, and gives no special privilege to the particular line of noise that a transcriber might subsequently be interested in. (Users of hearing aids often complain of a similar problem.) Nevertheless, remnants of the cocktail party effect can help the transcriber: concentrated listening to a sufficiently long stretch of speech can enable the transcriber's perceptual mechanism to tune in to the attended speech, and push even quite loud sound into the perceptual background.

We have seen above that when speech is heard in noise, the perceptual system relies even more heavily than usual on expectations and predictions based on knowledge of the language, knowledge of the topic or situation, and understanding of the acoustic context of speech – and that perception can be influenced by any of these factors, even more so than is the case with clear speech. When transcribing difficult material the transcriber has (or at least should have, as we will emphasize below) no knowledge of the context or situation. However the speech in the recording itself can create a linguistic context which is highly useful to the transcriber.

The second factor, then, is the duration of the recording. It must be long enough to provide such linguistic context. A related consideration is the consistency of the recording – an ongoing conversation or set of conversations among a small set of people is much easier to work with than a set of isolated utterances.

Factors relating to the transcriber

We have seen that most people, in perceiving speech, especially under noisy conditions, are highly influenced by their top-down expectations about the message based on knowledge of the context in which it occurs. This means that the transcriber of material to be used in a court case must be absolutely independent of, and disinterested in, the case, with no knowledge of the circumstances in which the transcription was made, or the external facts surrounding the case, such as names, events, places, etc. This is to avoid any suggestion that their interpretation was tainted by even subconscious preconceptions about what the people were likely to be talking about.

An advantage of the transcriber having no prior knowledge of the case is that any information he or she finds in the recording can be checked against the real facts of the case to provide a test of the accuracy of the transcription. If a transcriber with no knowledge of the case hears linguistically unpredictable information such as names and events discussed in

the recording, and these match those of the actual case, this suggests that the transcription has a certain level of accuracy.

We have also seen that people are generally very uncritical of their interpretation of speech, simply accepting their initial perception as correct; and that where perceptual errors occur, they are patterned, not random, showing reliance on stressed syllables and rhythm, well-known patterns of confusability of sounds, etc.

For these reasons, transcription of difficult material must be done by someone who understands the likely errors in their perception, and can be critical of their own perception, so as to consider a wider range of options, and judge the level of confidence to assign to any interpretation. Such knowledge is only gained through considerable study of linguistic phonetics and psycholinguistics.

This means that the transcriber of disputable recordings must be thoroughly trained – to the level of a postgraduate qualification – specifically in linguistic phonetics and psycholinguistics. This is because, as we have been at pains to emphasize, effective transcription depends upon understanding certain counterintuitive facts about the nature of speech, which is actually quite different from what most people think. Even many speech and audio specialists, such as speech pathologists, language teachers and audio engineers, unless they have specific, additional training in phonetics and psycholinguistics, often hold these inaccurate beliefs, as we saw in relation to phonemic transcription above. This can cause them to make gaffes like the following. In a case I was involved in, a speech engineer asserted that the phrase (from an advertizing jingle) 'It's Mac time' could be converted to 'It's smack time' by 'cutting out the space between the words'. Of course in normal speech there is no 'space between the words'; they flow into one another. Changing 'It's Mac time' into 'It's smack time' involves adding acoustic material, not deleting a space.

What is really happening in cases like this is that the sound engineer – who is no doubt an expert in his own field but not in phonetics – is treating speech as simply another kind of sound. Certainly speech *is* a kind of sound, and for many purposes, it is sufficient to treat it as such – for example for telecommunications, or for audio engineering, and even for some branches of phonetics. However, speech is not *just* another kind of sound; it is a very distinct kind of sound – sound that represents language. For certain purposes it is essential to understand the special nature of *linguistic* sound – and transcription of difficult material is one of those purposes. It is for these purposes that a background in linguistic phonetics (not just in the acoustics of speech) is essential.

It is common for speech specialists without sufficient background in linguistic phonetics and psycholinguistics to put too great a trust in the acoustics of speech as the determiner of phonemic perception. With appropriate equipment and training, it is possible to obtain visual repre-

sentations of the acoustic structure of speech – as waveforms, spectrograms, and a host of other representations, some very complex. One can measure, with great accuracy, a large number of acoustic characteristics of speech, and perform sophisticated statistical analyses on these measurements. This can be useful for various purposes but a problem potentially arises, however, when these measurements are taken as determinative of words or phonemes in a transcript. As we saw from the handwriting analogy above, no amount of analysis of one phoneme alone can fully determine which phoneme the speaker intended to produce. Acoustic analysis can certainly aid in arguing for one interpretation of the speaker's intention over another, but any such argument must always refer also to the wider context in which the individual sounds occur. This is why, as we have already argued, reliable transcription depends upon having a sufficient duration of speech for this wider context to be available.

The point is that to fully understand speech and speech perception – an essential background for transcribing difficult recordings and a range of other tasks – it is necessary to understand not just the phonemic or linguistic structure of speech, and not just the acoustic structure of speech, and not even *both* the linguistic and the acoustic – but the *relationship* between the two. This requires education to the level of postgraduate qualifications specifically in linguistic phonetics and psycholinguistics.

USING TRANSCRIPTS IN COURT

So far we have talked mainly about creating transcripts and judging the accuracy of existing transcripts. However, what we have said has important implications for how transcripts are used in court. Often it is a jury or magistrate who will ultimately decide upon the reliability of a transcript in relation to a particular case. The manner in which the transcript is presented to the court can have serious implications for their ability to make this decision. This is because their perception of the recording can be 'contaminated' by their prior expectations about its content.

Ideally a transcript should not be presented to the court until it has been agreed by both sides. Of course, in many cases this is not possible, since it is the court who will have to arbitrate the agreement. But every care should be taken to ensure that recordings and alternative transcripts are presented in a manner which maximizes their ability to reach a fair conclusion as to its reliability.

Finally, it should be recalled that simply having an accurate transcript of what someone said is often not enough to be sure of their meaning or intention in saying it. For example, providing evidence that someone used the word 'heroin' is not enough to show that they were talking about drugs. Examination of the wider linguistic context is necessary to ensure that they were not, for example, discussing the heroine of a film. Further, even if the linguistic context shows they were talking about drugs, were

they doing so in an incriminating way? Perhaps they were merely discussing the news, or joking around. The work of Roger Shuy is an essential reference for these 'higher level' issues in conducting a legal case based on transcript evidence (Shuy 1993, 1998).

RECOMMENDATIONS

Several conclusions can be drawn from the discussion so far, and framed as recommendations for making and using transcripts as evidence in court, or for judging the reliability of existing transcripts. These recommendations should be interpreted in the context of the foregoing discussion.

1. Transcription of disputable material should not be done by someone who has, or could be seen to have, an interest in the interpretation of the speech. If it is necessary for police to make their own transcripts of disputable material, these should be thoroughly checked by an independent expert before presentation to a court.
2. Where a recording is so poor that the transcript is likely to be contested, transcription should be done by someone with expertise in linguistics and phonetics, ie. postgraduate qualifications sufficient to impart a critical understanding of key references in forensic phonetics such as Baldwin and French (1990), Gibbons (2003), Hollien (1990), Levi and Walker (1990), Nolan (983, 199), Rose (2003).
3. Where there are disputes over particular words or phrases within a transcript, these cannot generally be resolved simply through acoustic analysis of individual disputed sounds, but require reference to the context available in longer stretches of speech, and careful linguistic and phonetic analysis by a suitably qualified expert.
4. A transcript of disputable material should show the transcriber's levels of confidence, and/or possible alternative interpretations, for each part of the transcript.
5. Where confidence is very low, for example where a recording is both of poor quality and very short, or a disputed word or phrase is isolated from its context (e.g. through extreme background noise, tampering, or equipment failure), or the overall quality is not uniformly poor, but shows inconsistency in recording conditions, material is generally best declared untranscribable.
6. No transcript should be presented to a jury before it has been vetted for accuracy by a suitably qualified expert in relevant branches of linguistics. This is to avoid 'contaminating' the jury by giving them expectations about the speech they are about to hear. Presentation of audio material and transcripts should follow the recommendations of Shuy (1998).
7. In general, though this is of course up to each individual court to decide, evidence from transcription of a poor quality recording is not sufficient to make the sole ground for a case, and should be used only in conjunction with other evidence.

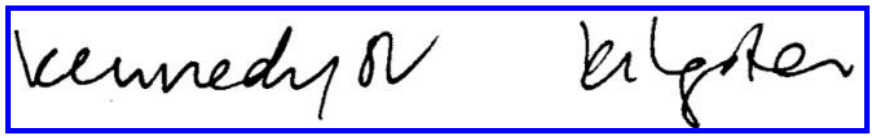

Figure 2 The context from which the excerpts in Figure 1 were excised

The text reads 'Kennedy St Kingston'. See the section 'Continuous not discrete' for discussion.

ACKNOWLEDGMENTS

An earlier version of this article was prepared for the New South Wales Director of Public Prosecutions in April 2001. The author thanks the editors for helpful suggestions in redrafting the article for publication and also Dr Phil Rose for useful comments and discussion.

REFERENCES

Baldwin, P. and French, P. (1990) *Forensic Phonetics*, London: Pinter.

Berko-Gleason, J. and Bernstein Ratner, N. (eds) (1993) *Psycholinguistics* (2nd edn), Fort Worth: Harcourt Brace Jovanovich.

Bond, Z. (1999) *Slips of the Ear*, San Diego: Academic Press.

Brown, G. (1977) *Listening to Spoken English*, London: Longman.

Bruce, D.J. (1958) 'The effect of listeners' anticipations on the intelligibility of heard speech', *Language and Speech*, 1: 79–97.

Byrne, B. (1998) *The Foundation of Literacy: The Child's Acquisition of the Alphabetic Principle*, Hove: Psychology Press.

Carney, E. (1997) *English Spelling*, London: Routledge.

Clark, J. and Yallop, C. (1990) *An Introduction to Phonetics and Phonology*, Oxford: Blackwell.

Eades, D. (1996) 'Verbatim courtroom transcripts and discourse analysis' in H. Kniffka (ed.) *Recent Developments in Forensic Linguistics*, Frankfurt: Peter Lang, 241–54.

Fillenbaum, S. (1971) 'Psycholinguistics', *Annual Review of Psychology*, 22: 251–308.

Fraser, H. (2001) *Teaching Pronunciation: A Guide for Teachers of English as a Second Language (CD-ROM)*, Canberra: DETYA (Distributed by Language Australia).

Fromkin, V. (1973) *Speech Errors as Linguistic Evidence*, The Hague: Mouton.

Fromkin, V. (ed.) (1980) *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*, San Francisco: Academic Press.

Gibbons, J. (2003) *Forensic Linguistics: An Introduction to Language in the Justice System*, Oxford: Blackwell.

Haberland, H. (1994) 'Written and spoken language: relationship' in R. Asher (ed.) *Encyclopedia of Language and Linguistics*, Oxford: Pergamon.

Hollien, H. (1990) *The Acoustics of Crime: The New Science of Forensic Phonetics*, New York: Plenum Press.

Just, M. and Carpenter, P. (1987) *The Psychology of Reading and Language Comprehension*, Boston: Allyn and Bacon.

Ladefoged, P. (1993) *A Course in Phonetics* (3rd edn), Fort Worth: Harcourt Brace Jovanovich.

Levi, J.N. and Walker, A.G. (eds) (1990) *Language in the Judicial Process*, New York: Plenum Press.

Miller, G.A. (1954) *Language and Communication*, New York: McGraw-Hill.

Miller, G.A. and Nicely, P.E. (1955) 'An analysis of perceptual confusions among some English consonants', *Journal of the Acoustical Society of America*, 27: 338–52.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition*, Cambridge: Cambridge University Press.

Nolan, F. (1997) 'Speaker recognition and forensic phonetics' in W. Hardcastle and J. Laver (eds) *A Handbook of Phonetic Science*, Oxford: Basil Blackwell, 744–67.

Olson, D., Torrance, N. and Hildyard, A. (eds) (1985) *Literacy, Language and Learning: The Nature and Consequences of Reading and Writing*, Cambridge: Cambridge University Press.

Osgood, C. and Sebeok, T. (eds) (1965) *Psycholinguistics: A Survey of Theory and Research Problems*, Bloomington: University of Indiana Press.

Pollack, I. and Pickett, J. (1963) 'The intelligibility of excerpts from conversation', *Language and Speech*, 6: 151–71.

Rose, P. (2003) *Forensic Speaker Identification*, London: Taylor and Francis.

Saporta, S. (ed) (1961) *Psycholinguistics: A Book of Readings*, New York: Holt, Rinehart and Winston.

Shockey, L. (2003) *Sound Patterns of Spoken English*, Oxford: Blackwell.

Shuy, R.W. (1993) *Language Crimes: The Use and Abuse of Language Eevidence in the Courtroom*, Oxford: Blackwell.

Shuy, R.W. (1998) *The Language of Confession, Interrogation, and Deception*, Thousand Oaks, CA: Sage Publications.

Spencer, A. (1996) *Phonology: Theory and Description*, Oxford: Blackwell.

Tent, J. and Clark, J.E. (1980) 'An experimental investigation into the perception of slips of the tongue', *Journal of Phonetics*, 8: 317–25.

Venezky, R. (1970) *The Structure of English Orthography*, The Hague: Mouton.

Walker, A.G. (1990) 'Language at work in the law' in J. Levi and A.G. Walker (eds) *Language in the Judicial Process*, New York: Plenum Press, 203–44.

Wray, A., Trott, K. and Bloomer, A. (1998) 'Transcribing speech orthographically' in A. Wray, K. Trott and A. Bloomer, *Projects in Linguistics*, London: Arnold, 201–12.