TOWARDS A COGNITIVE MODEL OF PHONETIC TRANSCRIPTION

Rachael-Anne Knight

City University London
knight@city.ac.uk

ABSTRACT
Phonetic transcription is crucially important to those who use it since it allows the ephemeral speech signal to be fixed for later analysis. However, we currently know little about which cognitive processes underlie transcription, or how transcription might be modeled. This paper asks if, in principle, cognitive models of language processing can be extended to incorporate transcription. It suggests a provisional model of transcription, describes how errors in transcription might provide evidence for the model, and discusses implications for pedagogy.

Keywords: phonetic transcription, language processing, cognitive models.

1. INTRODUCTION
Transcription involves noting down the sounds of speech using a special set of symbols, usually those from the International Phonetic Alphabet (IPA) and, provides a means to fix the ephemeral speech signal for later examination and analysis. Whilst transcription is sometimes aided by recordings and speech analysis packages, many professionals still transcribe live material without these devices, and it is this skill of unsupported transcription which we explore here.

Currently, we know rather little about the processes underlying phonetic transcription, which, as a complex cognitive-motor activity, is worthy of further study. This paper asks if, and how, a cognitive model of language processing can be extended to incorporate transcription. Such a model would deconstruct transcription to provide a greater understanding of the underlying processes. In turn, a model will provide a window into language processing in general, and may help educators give tailored feedback to students, and to design assessments which tap different underlying processes.

As a first step in modeling the transcription process we can simply introspect about the stages a person goes through when they transcribe. Firstly, they must perceive the signal in some way, usually through hearing, although vision may play an important role. All or part of the signal must be stored in short-term memory and a plausible segmentation made. Each putative segment in memory must be matched to stored representations of sounds, one of which must be chosen, converted to the appropriate symbol and written down.

2. MODELING TRANSCRIPTION
2.1. Models of language processing
The next step in modeling is to see how far these stages of transcribing can be accounted for by an existing model of language processing. The reasons for trying to incorporate transcription into an existing language model are twofold. Firstly, whilst transcription is a late learnt and specialist skill, it seems sensible to relate it to other aspects of language processing, especially as transcription may be subject to influences from reading and spelling. The second reason is one of parsimony. It seems unnecessary to propose a specialist model of transcription if extant models can provide suitable underlying functions, and we might expect existing models to contain appropriate architecture since transcription is simply another way in which humans process language.

Models of speech and language processing have a long history, ([1] provides a useful review), and today there are a number which aim to model different aspects of language. We will begin by attempting to incorporate transcription into the Ellis and Young model (e.g. [2]) as this is one of the most popular, and includes a variety of language processing pathways. We begin here with an information processing model as a first step in understanding transcription, in the hope that later work might allow computer modeling, and we also concentrate on the segmental level.

The model is shown below in Figure 1. Original modules from [2] are shown as red
rectangles, whilst the grey-filled ovals have been added to accommodate transcription, as we will see shortly.

**Figure 1: Proposed model of phonetic transcription based on [2].**

The original model covers single word production and perception in spoken and written media. Stimuli are detected by hearing or vision and are subject to various processes before being output in the desired form. For example, to spell a new or nonword, the acoustic signal is subject to ‘auditory analysis’, which divides it into segments in order to form the ‘phoneme level’. Each phoneme is then converted into a grapheme, using conversion rules based on those that are most common in the language, in order to form the ‘grapheme level’. Each grapheme is converted into the appropriate allograph (which includes choosing between a capital and lowercase letter), and motor patterns form each allograph to produce writing.

How far then, can this model be used for transcription? On the face of it a number of crucial elements are already in place. For example, the model has the necessary input and output modalities (hearing and writing), as well as modules for analyzing the speech signal. So, initially it seems that the model might work well, yet a number of elements need to be added or amended, as we will now explore.

### 2.2. A specialized transcription route

The first set of additions amount to adding an entirely new transcription route to sit alongside the model’s reading/writing and hearing/speaking routes, as shown on the right of Figure 1. Clearly, no general language processing model could be expected to included such a route, since transcription is learnt late, and only by a small percentage of people. The transcription route will require a number of specialist modules which are added during phonetic training and further refined with practice and experience.

#### 2.2.1. IPA sound representations

Within the specialist transcription route there is a need for a store of sound representations to which incoming stimuli can be matched. Some sounds will be available from the native language, but must be reanalyzed so that subphonemic differences can be perceived, whilst non-native sounds must be learnt from ear-training classes.

Currently there is little evidence about the form of these stored representations (which we will refer to as ‘IPA representations’). One area of interest relates to whether they are stored in an episodic fashion, as appears to be the case for real words ([3]) and native speech sounds ([4]), or if they are more abstract. Studies of second language learning (e.g. [5]) have shown greater robustness of mental representations for newly learned speech sounds if there is variation in the voices heard during training. This suggests that such representations may indeed be episodic. However, learning the sounds of another language is rather different to learning non-native sounds outside the language learning context. In phonetics ear-training classes, sounds may be pronounced in isolation, explicit instruction may be given about pronunciation, and there will rarely be a meaningful linguistic context. At present then the form of the IPA representations is unclear, but we will assume a module where such representations are stored, in order to be matched against incoming signals.

#### 2.2.2. Symbol store

There must also be a store for the symbols that represent speech sounds. Learning new non-iconic symbols is rather unusual in adult life, and we know little about how they are acquired. We do know, however, that non-iconic symbols (such as Blissymbols) require more presentations to be learnt, and are retained less readily, than iconic symbols, at least by adults with physical disabilities ([6]). The symbol store might initially be populated by symbols which have a similar value in the IPA and the orthography of the transcriber’s native language, such as [p b s] for English speakers, whilst others will be learnt by careful practice and usage. The symbol store is not shown explicitly in the model but is incorporated
2.2.3. Output

The transcription route’s output processes are likely to rely heavily on those used for writing, so separate modules for transcription are not proposed. As many IPA symbols are similar to letters from other alphabets, it seems likely that transcribers rely on their orthographic knowledge and use similar graphemic motor patterns for transcription, but with some additions and modifications for less familiar symbols.

2.3. Details of auditory analysis

In addition to a specialized route for transcription, other aspects of the Ellis and Young model need modification in order to account for transcription. In particular, more detail is needed about the auditory analysis that is applied to an incoming signal. In the original model, auditory analysis simply involves segmenting the speech signal and deciding which phonemes it contains. These phonemes will then be represented at the ‘phoneme level’, where they can be turned into speech, or converted into graphemes for spelling.

The major issue in adapting the model for transcription is that incoming sounds will not always be phonemes of the native language, and being able to hear subphonemic differences and process non-native sounds, is crucially important. Here the model of Stackhouse and Wells [7] provides a useful counterpoint to that of Ellis and Young. Their psycholinguistic framework distinguishes between ‘phonological recognition’ (roughly the equivalent of the ‘phoneme level’ in Ellis and Young’s model) and ‘phonetic discrimination’. ‘Phonological recognition’ occurs when speech input is recognized as native and is passed on to the mental lexicon for further processing. ‘Phonetic discrimination’ is considered to be an offline process, used when listening to an unfamiliar accent, processing paralinguistic sounds, and learning a second language. For many tasks then transcribers will have to use ‘phonetic discrimination’ rather than ‘phonological recognition’, as they must focus on the full range of IPA sounds. ‘Phonological recognition’ may be sufficient for systematic transcriptions of the native language, but impressionistic transcriptions will require phonetic discrimination. Hence, our model must include modules for both types of input processing, as shown towards the left of figure 1.

Another aspect of auditory analysis which is not provided by the Ellis and Young model is a module for short-term memory, which has been implicated in transcription by [8]. Currently the most influential model of short-term memory is that of [9] (and subsequent revisions). Here incoming auditory representations are stored in a phonological loop, where they decay over time, but can be refreshed by sub-vocal rehearsal. The loop is limited, either in time to around a couple of seconds, or in slots to around seven items.

The model of [10] includes a ‘phonological buffer’ which can be considered as the equivalent of the phonological loop and is thus added to our model of transcription after ‘auditory analysis’ in order to reproduce the effect of temporary memory storage on transcription.

2.4. Information flow for transcription

The proposed modifications to the Ellis and Young model allow transcription to be modeled as follows: In order to transcribe a heard stimulus, the signal is segmented by the ‘auditory analysis’ system, and then passed to the ‘phonological buffer’ for short-term storage. The route taken then depends on the type of transcription required. For a systematic transcription of the transcriber’s native language the signal is passed to the ‘phonological recognition’ module, whereas for impressionistic transcriptions the material is passed to the ‘phonetic discrimination’ module. Both modules feed into the ‘phoneme level’ (which would more accurately be termed a ‘segment level’) where each segment in the stimulus is specified. From here the segments are matched against ‘IPA sound representations’ to identify the appropriate sound. Then the relevant symbol is accessed (via ‘sound-symbol conversion’) and written down using modified graphemic motor patterns.

It is likely that the entire process is iterative, with various hypotheses considered at each stage, so that the transcriber can think back over the material stored in the buffer (which can be refreshed with subsequent repetitions and sub-vocal rehearsal) in order to fine-tune their transcription. It is also possible to transcribe from orthography via the ‘visual analysis system’ or from internal speech via the ‘semantic system’,
each of which pass information to the ‘speech output lexicon’ and in turn to the ‘phoneme level’ before entering the transcription route.

3. ACCOUNTING FOR ERRORS

One way to test a cognitive model is to determine if it can explain errors in human performance. For example, [2] uses errors such as over-regularization as support for a non-lexical spelling route. This section briefly suggests how two common errors made by student transcribers might be explained by the model.

Errors can occur at any level of the transcription model, and those which occur at some levels represent a more severe problem with knowledge and understanding than errors which occur at others. Consider two students, both presented with [ŋ] in a transcription task. One believes they have heard a velar nasal and thus transcribes [ŋ]. The second correctly identifies the sound as a retroflex nasal, but erroneously writes the symbol [ŋ]. Both students have given the same wrong answer, but the errors occur at different stages in the model, and the educator would likely be more concerned about the first error in auditory analysis, than the second, more peripheral error in sound-symbol conversion.

Whilst educators will likely comment on such points in class, many assessment practices make it impossible to differentiate between the loci of these errors, as only the finished transcription is available for analysis. One implication is to design assessment tasks which differentially tap each level of the transcription model so that students can be given feedback tailored to their needs. For the example above, a comparison of the results from carefully designed analytic listening tests ([11]), and label to symbol matching tasks would allow the educator to assess students’ relative strengths and weaknesses.

Additional errors result from interference from other modules within the language processing system. For example, novice students, transcribing from orthography, often transcribe the English word ‘I’ as /ɪ/. Here the symbol used in spelling is passed from the ‘visual analysis system’ directly to the ‘grapheme level’. The vowel sound is not passed to the transcription route and auditory analysis is not carried out. Such errors likely occur because reading and writing are highly automatic processes and difficult to suppress (as shown by the stroop effect). Thus moving from a novice to an expert transcriber ([12]) might be modeled as the addition of specialist modules (such as stores of IPA sounds and symbols) plus the suppression of other language processing routes, such as reading and spelling. Hence, tasks used in class and in assessments should help students to suppress more automatic processing routes. For example, some institutions use a task whereby students read aloud a transcription that looks like an English word (e.g. [coral]) but is pronounced very differently. This task helps students to suppress their natural tendency to read, and instead encourages them to interpret each symbol with its value in the IPA.

4. CONCLUSIONS

This paper has suggested a preliminary model, based on a pre-existing language processing model, which deconstructs the processes underlying transcription. In doing so, we might be able to model the levels at which student errors occur and tailor our advice to individual students accordingly. The model can also help us to think about how students learn transcription and to incorporate tasks into teaching and assessment that aid their transition from novice to expert.

5. REFERENCES

[1] www.smithsrica.demon.co.uk/PSTranscoding.html