Excerpt from

Tatham, M., and Morton, K. (2003). *Expression in Speech: Analysis and Synthesis*. Oxford: Oxford University Press.

INTRODUCTION

When people speak to each other they are able to communicate subtle nuances of expression. Everybody does this no matter how young or old, or which language they are speaking: the existence of expression as an integral part of how speech is spoken is completely universal. This does not mean that every language or speaker expresses everything in exactly the same way: they do not.

The expression people bring into their conversation often says something about their feelings about the person they are talking to, or perhaps something about how they feel about what they are saying, or even how they feel in general today. This expression is incorporated in what a person says without changing the words being used or the way these are arranged into sentences. Different expressions are conveyed by changes in the acoustic signal – using different 'tones-of-voice' – rather than by altering lexical choice (which words are being used) or sentential syntax (how those words are arranged in the utterance). Of course this does not mean that people *never* alter the words they using deliberately to convey expression: the point is that you can inform your listener directly about how you feel with words, or you can convey expression by a change of tone-of-voice. Tone-of-voice has expressive force, and is a very powerful means of telling people things the words themselves sometimes do not convey very well: what our attitude is and how we feel.

A consequence of the universality of tone-of-voice is that we never speak without it. We can imagine a kind of 'neutral' speech completely devoid of expression, but in practice it is safe to say this never actually occurs. Many researchers feel that a description of what neutral speech *would* be like is a good starting point for talking about different types of expression, but we shall see that this is likely to be an abstraction rather than anything which can actually be measured in an acoustics laboratory or deduced from people's perception. In real conversation anything we might call 'neutral speech' is speech with minimal or ambiguous expressive content, but it is not speech with *no* expressive content. In fact such speech would be extremely difficult to characterise precisely because it *does* contain expression but only in a minimally detectable way.

Listeners respond remarkably consistently to differing tones-of-voice. This means that however subtle some of the effects are they are part of our communicative system. If speakers regularly produce recognisable expressive tones-of-voice it follows that, at least at first glance, we should be able to detect in the speech signal differences which correlate well with listeners' feelings about expression. This is a very simple concept, but one which still largely defeats us. Tone-of-voice is apparently consistent for both speaker and listener, yet it remains quite elusive when we try to say something about what it is and how it works. It is part of the way in which we externalise our internal word using language.

Adding expressiveness to speech

It is usual to think of the speech signal – the acoustic manifestation of an utterance – as being the result of a complex chain of events which begins with some 'idea' or 'piece of information' the speaker wants to communicate. Most schools of linguistics feel that language is essentially a kind of encoding system which enables speakers' intentions to be turned into speech signals for 'transmission' to a receiving audience. Some of the encoding processes will be sequential, others will take place in parallel; and one of our concerns will be to decide at what stage expressive content comes into play.

Some of the properties of expressive tone-of-voice may be phonological in origin, associated with the planning of speech, before it is phonetically rendered or turned into an acoustic signal. Others may be phonetic in origin or associated with the control of how the phonological plan is rendered phonetically. How and where expression gets into the speech signal is very important because this can determine how much control a speaker has over the expression. Extreme examples might be when a strong feeling of anger can make a speaker's voice tremble in an involuntary way, or when a speaker subtly injects a tone of irony or reproach into their speech.

The basic model we shall be suggesting consists of a planning stage followed by a rendering stage. These two stages correspond roughly to the levels of phonology and phonetics in linguistics. An important refinement of this basic approach is to add that the rendering stage is closely monitored and kept accurate as rendering unfolds, a process known as 'supervision'. Rendering is a relatively new term in speech production theory corresponding to the earlier and simpler term 'realisation'. We shall be developing this general concept and the concept of supervision as we proceed, but will leave more complete definitions till later.

A computationally oriented approach

Because we intend suggesting an *explicit* model of expression in speech we have chosen to make that model computational – that is, capable of being computed. This simply means that our descriptions are of algorithmic processes which have a beginning and an ending, and clearly defined linking stages. The purpose of developing a computational model is that it will run on a computer. Since it is based on experimental observation the model is descriptive of how human beings have been speaking with expression, but at the same time the model is formulated to be predictive. In particular it predicts an acoustic speech signal incorporating expression, and we have chosen to set up the model in the test environment of *speech synthesis*. That is, we describe expression in speech in terms of how a soundwave might be created *synthetically* to simulate the speech of a human being.

There are problems with using a speech synthesiser to test models of human speech production and perception. One of the main difficulties is that the speech production and perception constraints at all levels have to be programmed in detail. In one sense this is a serious problem because it means we have to know a considerable amount about how human speech production works and the constraints which operate on it. Perhaps surprisingly, the existence of this problem is precisely why computational modelling is so important and so revealing: gaps in our knowledge, inadequate basic assumptions and shortcomings in our descriptions are clearly brought out, forcing careful consideration of every detail.

If our model of human speech production is descriptively and computationally adequate from the point of view of characterising what we observe in the human being's behaviour, it follows that its implementation as a synthesis device (either for testing or for more general application) will take us as close as possible to a simulation of the human processes. This does not mean however that the model tries to be indistinguishable from what a human being actually does. On the contrary, this is paradoxically what we do *not* want. A model which *is* what is being modelled is not a model at all, and therefore not able to fulfil the purpose of models: to cast light on the nature of the object being modelled. Such a model would have the same black-box characteristics as the object itself.

Speech synthesis

Because the computational model being used to create synthetic speech is based on our understanding of human speech production, the simulation incorporates the human properties which have been addressed. For example, in characterising human speech it is appropriate to distinguish between a phonological planning stage and a phonetic rendering stage. This distinction gets transferred to the simulation: phonological planning is treated as something we call 'high level synthesis', concerned with simulating cognitive processes in speech production, whereas phonetic rendering is treated for the most part as 'low level synthesis'. Low level processes are more about the physical production of the soundwave using descriptive models of the acoustic structure of speech.

The idea of high level and low level synthesis will be developed as we proceed with describing synthesising expression in speech. But we might notice here that there is important and revealing interplay between the two levels. For example, if we are satisfied that we have a good model of how stops coarticulate with the vowels which immediately following them in syllables, we are in a good position to model the high level plan which will enable the coarticulatory model to produce a good soundwave. Coarticulation theory models how the results of interacting segments are revealed in the linear stream of speech; a good model will predict the kind of input needed for this to happen. That is, how we model coarticulation interacts with how we model the parts of the higher level plan which are eventually to take part in the coarticulatory process. We use coarticulation as an example, but it is worth remembering that segments need a theory of how they coarticulate only if the overall model assumes that there *are* segments. The same principle applies to a model of expression: we need to establish a suitable construct about what we *mean* by expression – something we shall consider when we discuss in detail a possible model of prosodically based speech production which incorporates expression.

A computational model of expressive content in speech

Since part of the research community's purpose in modelling the expressive content of speech is to generate synthetic speech for both practical purposes and for the purpose of testing the model of human behaviour, we consider that a computational model is not only desirable but is *essential*. The model should have coherence and integrity by reason of the fact that it *is* computational, and should easily be able to be incorporated into high level aspects of speech synthesisers. But we need to ask some questions:

- Is a computational model appropriate for dealing with the phenomenon of expression in speech?
- Does the concept of expression lend itself to being computationally defined?
- Can the acoustic correlates of expression be determined and stated in a way which feeds into the computational nature of the model?

If we feel that computational modelling of expression in general and the expressive content of the acoustic signal in particular is both appropriate and possible, the next step is to attempt to determine physiological, cognitive (including social) and language-based contributions to the overall production of the speech signal.

An integrated theory of speech production and perception

An integrated theory of speech production/perception is rarely broached in the literature even in connection with simple segmental rendering. But it is even more rare to find discussion of full theory of prosodics (including expression) which integrates production and perception. Under these circumstances it would be reckless to attempt more than collate observations and try to begin the process of building a theoretical framework for modelling expressive content in speech. Bearing in mind that here there can be no last word claims, and that the best we can achieve is a statement coherent enough to be demolished, we shall discuss the preliminaries to a speech production model (Part IV, Chapter 4) which integrates both speaker and listener perspectives on the one hand and segmental and prosodic perspectives on the other.

Although we reiterate the point, let us repeat that this is no more than a useful working model which attempts to make sense of observations about human speech production and perception. Pending data to the contrary there is the weakest of hypotheses that human beings actually work this way. We stress again, though, that the model is not and cannot claim to *be* the human being: it is just a device of the scientist. Separate models of production and perception fail to account for apparent interdependence between the two modalities; we propose that an integrated model takes the theory forward in the sense that it pulls in the observations the other models neglect.

Our account of modelling expression

Part I is a general treatment of expression in speech. We examine how researchers have been treating the subject in investigating natural speech and transferring their findings to speech synthesis. We include a section of how expression is perceived by listeners, in particular discussing the non-linearity of the relationship between the acoustic signal and perception. We conclude that perception is an act of assignment rather than a process of discovery. Part II moves on to the detail of how researchers have transferred ideas about natural expression to the domain of speech synthesis, and include a discussion of recent developments of the technology. Our characterisation of synthesis involves a formal separation of high and low levels, corresponding to cognitive and primarily physical processes in human speech production.

Part III is devoted to an appraisal of the background research. Several disciplines are involved and how they come together is critical if we are to have a comprehensive understanding of expression. We begin with the biological and psychological perspectives, and move on to the linguistics, phonology and phonetics perspectives. We then examine what these approaches mean for speech synthesis and how they might point to a way forward. We examine how models of expression might be evaluated in the light of synthesis requirements.

Our concluding section, Part IV, begins by outlining the beginnings of a general model of expression, with a view to proceeding to a fully integrated model based on the findings covered in Part III. We look at a way of formalising the data structures involved and discuss why this is crucial to an explicit model. We evaluate expressive synthesis in terms of longer term developments which make prosody and expression central to the model. The final chapter moves to proposals for a model of speech production anchored in prosody and expression – an almost complete reversal of the traditional approach in speech theory. Expression, exemplified in emotion, becomes central to the discussion and envelops the model. Speech is characterised as a carefully planned and supervised process operating within the dominating requirements of expression.