THE MEASUREMENT OF VOICE

Jeffery Pittam
Department of English
The University of Queensland

ABSTRACT - This paper discusses a range of convergent methodologies for the measurement of voice. It argues for an integrated approach to vocal measurement that conceptualises the voice as an important part of social interactions. This approach includes acoustic, perceptual and attributional measures. Two projects currently being undertaken by the author will be briefly introduced.

INTRODUCTION

This paper argues for an integrated approach to the study of voice, one that focuses on vocal communication rather than solely measuring the acoustic properties of voice or indeed the perceptual characteristics of voice. The term *voice* is used here to include all articulatory settings (Laver, 1980) whether used paralinguistically or non-linguistically; that is, the long-term aspects of vocal communication called by Laver *voice qualities* including tension settings, plus the dynamic elements of vocal communication such as pitch and loudness. Some researchers (including the present author) have also included temporal features such as speech rate.

It is becoming increasingly important that our work in speech science and technology is placed into a broader social frame. In large part ours is an applied field, with applications such as speaker identification systems, speech synthesis, many applications in the telecommunications field, automatic translation systems, text-to-speech applications in fields such as communications devices for the disabled, and in education, The use and acceptance of such applications relies more and more on our understanding not only of the acoustics of voice but our perceptions of voice and the attributions we make based on the voice. Such information may provide both accuracy and quality to all applications, as well as ensuring greater acceptance of those applications by the users.

If we are concerned with vocal communication, that is, the use of the voice in social interaction, all elements of our identity need to be taken into account in our study of voice. This will include aspects of social identity such as age, gender and ethnicity as well as personal identity, plus the encoding of emotion and attitude into the voice. Clearly, a detailed account of all such areas cannot be achieved in this paper. The reader is referred, therefore, to Pittam (1994) for a much more comprehensive account of vocal communication in social interaction. Here, I will propose the form of theoretical framework that may be useful in this research, but will start by outlining some of the principles I feel are needed to underpin our physical measurement of voice. I will not attempt here to deal with perceptual or attributional measures.

PRINCIPLES UNDERLYING THE MEASUREMENT OF VOICE

Essentially, any of the physical measures used by speech scientists to examine speech are potentially useful to voice also. This can be seen most easily in the fact that voice tends to be described using the same types of parameters as those used for speech. This is so whether we consider articulatory measurement, aerodynamic or acoustic. Thus, with articulatory measurement, researchers in the voice area rely on position of

the various vocal organs, whether voicing is present, and the type of sound produced. With acoustic measurement, a similar point can be made.

The major difference between voice and speech measurement at this level is that voice researchers are not overly interested in the speech sounds themselves although even these may provide us with useful measures of the voice. We are particularly interested in the measurement of long-term vocal features, those that extend over more than a single speech sound or segment. Given this, the measures used are themselves usually long-term. This does not necessarily mean we must ignore short-term measures such as those relating to individual speech sounds, but that, if used, they will need to be statistically manipulated to simulate longer time frames. Means, summations, ranges or variances of the physical measure used, for example, may need to be calculated.

Measures underlying the perceptual features of pitch and loudness, as well as temporal features have all been used. With the first of these, fundamental frequency and its derivatives, jitter, formant frequencies - suitably manipulated to represent longer time frames - have all been used, as has intensity. More complex measures are also frequently used: measures based on long-term spectral analysis and inverse filtering.

One point worth making here is that if we are concerned with placing voice into its social context, the physical measures we adopt must be ones that can be used validly with perceptual measures. Some of the electronic measures that have been related to the voice (including some of those by the present author) have been so reduced statistically that it is difficult to validate their use with perceptual measures. If we are concerned with our perceptions of the vocal signal, the physical measures we use must have some meaning for the listener. This is not to say, of course, that such measures should not be examined and incorporated into our applications if relevant, when the focus of our attention is on the production of the signal only.

AN INTEGRATED FRAMEWORK FOR CONVERGENT METHODOLOGIES

Elsewhere (Pittam, 1994), I have argued that the study of voice in social interaction requires a theoretical framework that may be viewed as four independent but interacting dimensions. The aims and assumptions of the theoretical perspectives that represent each dimension will be sufficiently different not to warrant collapsing them down into a single entity. The proposed framework, then, starts with the individual speaker and, by a series of theoretical dimensions, places that individual into contact with others and the society of which he or she is a part.

I will deal with each dimension in a little more detail later, but for the moment let me introduce the whole framework. At one level we need a phonetic model that centres on the individual's voice. Essentially, Laver's (1980) model provides this. It is worth noting, however, that by linking articulatory settings to acoustic measures of the sound wave, Laver is moving away from the individual speaker as such, to a physical phenomenon that others can perceive. This links it to the next dimension: one relating precisely the physical measurement of voice to others' perceptions. Scherer's (1978; Goldbeck & Scherer, 1988) Modified Brunswikian Lens Model is such a theory, and just as Laver provided a link with the second dimension, so Scherer moves beyond perception to inferences and attributions, relating this dimension to the third one. At the third level we need models that show us how events such as social interactions develop and proceed. This introduces the models of social exchange. Communication Accommodation Theory (Giles & Coupland, 1991) and Patterson's (1983) Sequential Functional Model of Nonverbal Behavior belong in this category. Finally, we need a

broader perspective on the whole situation showing that social interactions are not just about individuals communicating, but are social events influenced by the institutional and generally socio-structural nature of society. Once again, the two social exchange models provide a link to this fourth dimension by allowing for social influences on the interaction. The links between the dimensions provide one way of bringing this all together. Another way, and one that highlights the relevance to vocal communication, is to develop a taxonomy of vocal function based on that material. This is something else dealt with in Pittam (1994). While acknowledging that many of these concerns are not directly the domain of the speech scientist, some are, and all need to be taken into consideration when designing our applications.

Before moving on to look at some areas of specific interest to speech science and technology, let me review the current position in all the areas covered by this proposed framework. The first thing that needs to be said, and which prompted me to develop this theoretical framework in the first place, is that work on the voice has been conducted in a somewhat piecemeal way for decades. One of the things that influenced Laver's early work was the lack of consistency in phonetic description and the lack of standardisation of terminology in the whole area of voice.

The first dimension of the framework is concerned with a standard phonetic description incorporating a standardised terminology to enable us to understand which physical parameters are salient and how to measure them. Laver's (1980) model provides a phonetic description and a vocal profiling system that at one level provides the answer to this dimension. Laver's model, however, is essentially articulatory and in its applied form - the profiling system - requires considerable training by its adherents that brings with it additional problems. The profiling analysis is a perceptual system which requires the trained user to perceive settings of our articulatory organs and to scale them. This introduces concepts such as the need for accuracy of initial judgement, consistency across different users and reliability in terms of replication, as well as the need to develop norms for all vocal features measured for all the communities studied. This is a difficult task, albeit not an impossible one, that involves asking non-trivial social questions about the setting of the norms in the first place. Despite this, the profiling system represents a major step forward and is a method I would like to see used more frequently - and not just in the area of voice pathology, which prompted its initial development. Laver's model itself, as a phonetic description, potentially at least provides us with articulatory, acoustic and aerodynamic descriptions of all the articulatory settings that we use. Unfortunately, however, much of this information has still to be collected in a consistent, reliable way, and presented using standardised terminology enabling all researchers to know exactly which voice features are being discussed and allowing comparison of results.

The second dimension is concerned with linking elements of our identity or our psychological state to their physical manifestations, and linking both to listeners' perceptions and subsequent attributions made about the speaker. Scherer's (1978; Goldbeck & Scherer, 1988) Modified Brunswickian Lens Model provides the only means of theorising this at present. As I presented some thoughts on this at the last conference, I will not go into great detail here, but a number of points can usefully be made. To start with, little work has been carried out testing this model, even by Scherer. He has used it to examine the salient vocal parameters for personality characteristics, such as extroversion, and for some basic emotions. As is indicated below, I am about to start examining the vocal characteristics of ethnic identity and will be incorporating Scherer's model into this work, thus testing the theory and hopefully extending our knowledge of salient vocal features and their acoustic characteristics.

One of the problems with the model as it has been used to date is a propensity to use some non-standardised terms to describe the perceptual features of the emotions or personality characteristics. In fairness to Scherer, however, he does provide definitions of his terminology allowing us to equate it with other existing terms. There is little doubt in my mind that Scherer's lens model will provide the conceptual underpinning to all aspects of vocal communication that relate to this second dimension.

Before dealing briefly with the other two dimensions of the proposed framework, it is worth adding that there is at least one further problem relating to the first two dimensions. This concerns the functions of vocal features. I suggested above that a taxonomy of vocal function was one way to bind the framework together. I have attempted to develop such a taxonomy, and once again details can be found in Pittam (1994). This does not overcome the awesome problem of multifunctionality of vocal features nor the concurrent saliency of more than one function, however. Thus, any vocal feature may function in more than one way. It may, for example, be characteristic of a particular emotion, be a feature of age or gender identity - or both - as well as functioning linguistically. Intensity, fundamental frequency and temporal features have all been shown to serve linguistic functions and to communicate personality characteristics in addition to emotion. Similarly, aspects of identity, emotional state, and linguistic functions, to name but a few, may all be operating simultaneously. Perhaps the most pressing need in the voice area is to untangle these two problems. In addition, of course, presence or absence of a feature does not tell the whole story; many features such as frequency and intensity are always present. Degree of presence (as with intensity, for example) or bands within the total range of a feature (as with frequency) may also be important to differentiate how a particular vocal feature functions.

Some headway on such problems has been made. There is some evidence suggesting that linguistic phenomena and emotion are interrelated in the acoustic signal. Liberman (1978) has pointed to the slight variations in intonation that provide the listener with information about subtle shifts in affect and attitude, while Cahn (1990) has shown a relationship between linguistic stress (emphasis) and emotion. Evidence is also beginning to emerge (Hermansky and Cox, 1991) that linguistic and speaker dependent information can be separated during speech analysis. If so, it would be a significant step forward in our understanding of vocal communication.

I do not intend to say much about the remaining two dimensions of the framework, neither is as salient to speech science and technology as the first two dimensions. As already suggested, however, both are important and we certainly need to keep both in mind when engaged in our work - particularly if our work is in an applied area. Dimension three is concerned with models of social exchange to enable us to trace the interactional sequencing of vocal function. This actually provides us with yet another problem akin to the ones just raised: the dynamic nature of vocal function across the space of an interaction, even a single sentence. Little work has been conducted at this level. A few features, particularly ones relating to the management of interactions have been explored, but even here, the dynamic nature of vocal communication between interactants has been largely ignored. The two theoretical models proposed as likely candidates for any examination of this area have not been operationalised or tested for their applicability to vocal function.

Finally, the fourth dimension, the means of linking the interaction (and therefore vocal function) back to its social context, is not only the area most removed from our work,

but is also the one that has received the least attention. No theoretical model currently exists that addresses this in anything like a comprehensive way.

The Modified Brunswikian Lens Model

In the proposed framework, the second dimension - the one relating the physical measurement of voice to others' perceptions - is the one I want to say a little more about here. This is one of the dimensions that has most direct concern for speech science and technology in that it links the actual physical measurement of the voice to perceptual measures. It is also the one that I will be testing in the two projects to be introduced briefly below.

Having discussed this two years ago, I will simply introduce the basic conceptual underpinning of the model and its major components. The first part of Scherer's model statistically links psychological characteristics with their physical counterparts. Other statistical links are then made with a listener's perception of the physical feature and the attributions about the speaker that are subsequently made. One of the most important features of such models as this is the relationship between utilization and the ecological validity of the physical features. Ideally, one utilizes (perceives and makes inferences about) only those features that are ecologically valid (those features that really do stand for the things we take them to represent). The measurement of ecological validity is represented in the model by the correlation coefficient between the psychological characteristic and the physical measure of the vocal feature. Thus, Scherer has shown that extroversion is significantly correlated with a number of vocal features such as frequency, intensity and variation in frequency. These vocal features are perceived by the listener as pitch, loudness, and what Scherer calls sharpness and gloom - two of the non-standard terms mentioned earlier - and from these the listener infers that the speaker is extroverted. When using this model, the researcher hopes to find significant correlations between pairs of concepts. Such linkages enable the researcher to isolate the specific vocal features that communicate the trait.

Clearly, this model depends on the existence of a stable relationship between voice and, in this example, personality characteristics in the first place, as well as our ability to isolate relevant vocal cues. Given this, Scherer suggests, accuracy of judgment also depends on a large degree of correspondence between actual and inferred voice-personality correlations. The use of the lens model is in part an attempt by Scherer to overcome some of these problems.

CURRENT RESEARCH

Before concluding, let me add a brief note about two research projects in which I am currently involved - both very much in their infancy. The first project concerns the importance of key vocal features to the area of telemarketing. Currently it does not involve the physical measurement of the voice, only the perceptual measurement, so I will be brief. As an applied area, this has received little attention, yet it is potentially of great interest. This is one of the few areas in which voice can play a major role in determining the success or otherwise of the task - persuading people to buy. It does have, of course, other analogous applications in the telecommunications area - what works in live telemarketing is likely to be useful in other telecommunications domains requiring, say, speech synthesis. This is a project comparing the characteristics of the voices of speakers who are proven successes in telemarketing with those who are less successful. Unfortunately, as yet I have no results to present, and can only flag that the project is under way.

The other project I have already made reference to: the vocal communication of ethnic identity. This is even more in its infancy than the first. However, I will just say that in this project we are seeking to tease out the vocal characteristics of ethnic identity in two groups: Vietnamese Australians and Chinese Australians from Hong Kong. It is this project that will provide the opportunity to examine the problems of multifunctionality of voice features and simultaneity of functions, in addition to the central aim of highlighting those vocal characteristics salient for ethnic identity. Once again, this is an area little studied - particularly when we consider both physical measures and perceptual measures and the relationship between the two.

SUMMARY

In this paper, I have attempted to show the extent of the difficulties facing a researcher in vocal communication. The proposed theoretical framework is designed to provide the conceptual underpinning that may point the way to overcoming some of these difficulties. Above all, however, this paper tries to highlight the need to take account of more than just the physical measurement of voice; to accept that voice is an essential part of social interaction and that parts of the interaction other than the actual production and characteristics of the vocal signal must have a bearing on what we as speech scientists do.

REFERENCES

Cahn, J.E. (1990) "The Generation of Affect in Synthesized Speech", Journal of the American Voice I/O Society 8, 1-19.

Giles, H., & Coupland, N. (1991) *Language: Contexts and Consequences*, (Open University Press: Milton Keynes).

Goldbeck, T., Tolkmitt, F. & Scherer, K.R. (1988) "Experimental Studies on Vocal Communication", In K.R. Scherer (Ed.), *Facets of Emotion*, pp. 119-138, (Lawrence Erlbaum: Hillsdale, NJ).

Hermansky, H., & Cox, L.A. Jr. (1991) "Perceptual Linear Predictive (PLP) Analysis-resynthesis Technique", In G. Pirani (Ed.), *Proceedings of the Second European Conference on Speech Communication and Technology*, pp. 329-332, (ESCA: Genova).

Laver, J. (1980) *The Phonetic Description of Voice Quality*, (Cambridge University Press: Cambridge).

Liberman, M.V. (1978) *The Intonational System of English*, (Indiana University Linguistics Club: Bloomington).

Patterson, M.L. (1983) *Nonverbal Behavior: A Functional Perspective*, (Springer-Verlag: New York).

Pittam, J. (1994) *Voice in Social Interaction: An Interdisciplinary Approach*, (Sage: Newbury Park).

Scherer, K.R. (1978) "Personality Inference from Voice Quality: The Loud Voice of Extroversion", European J. Social Psychology 8, 467-487.