Proceedings of the Institute of Acoustics

SOME ISSUES IN INVESTIGATING SPEECH STYLES

Luke Boucher, Sheila Williams, Sandra Whiteside

Centre for Language Engineering, University of Sheffield, Sheffield, S10 2TN.

1. INTRODUCTION

This paper considers some of the issues surrounding the collection of a database contrasting the two styles of "Casual" and "Careful" speech. One of the aims of the VOX project*, of which this work is part, is to study speech style variation for use in Speech Synthesis, and the discussions in this paper will be made with reference to the data collection technique that has been developed for the project at LIMSI in Paris. Recording is well underway at LIMSI as part of the French language contribution to VOX, and at Sheffield we are implementing a variation of this technique, suitable for English, but with results that can be compared to the French.

The term, Speech Style is taken to refer to certain types of intra-speaker variation. Here we must distinguish between the variation due to the emotive state of the speaker, that due to linguistic repertoires associated with social and regional variation and that due to repertoires associated with situational context. It is only the third of these types of variation which we address here. That is, the different ways in which someone will pronounce the same words, for example, sometimes saying /did ju/ and at other times /dja/ or /diju/ for "did you". For instance, a casual style would be distinguished from the phonemic reductions associated with the informant being in a depressed state or very excited and a careful style would be distinguished from a formal register, in that formal speech belongs to a linguistic repertoire [1] which might have both words and syntax that differ from a non-formal register. Thus, the term 'style' is intended here to capture only variation due to different tasks, settings and audiences, and the way in which these interact with the speaker's perception of the audience requirements. Thus a speaker, while using a particular dialect, may enunciate it in a different manner according to the perceived difficulties of the communicational setting or the attributed importance of accuracy in conveying the message.

Amongst researchers working on speech style there may not be complete agreement over the way in which it has been defined in this paper. The term does not have a stable technical definition as yet [2] and in this work we are adopting, what might be seen as, a rather restrictive view. It may be more precise than other uses but this is not without problems; for data collection purposes, the definition must relate also to what it is possible to record and measure, emotive factors can not really be excluded so easily from different recording conditions. Definitions of Style, and the terms used to refer to different styles, are also dependent upon the psychological theory underlying explanations of why these variations occur, and some of the existing theories will be presented in this paper. The paper will also look at practical attempts to study style contrasts and present the method for style data collection being applied at Sheffield.

There is, then, much scope for discussion and it is hoped that this paper will make a useful contribution.

* VOX Working Group: ESPRIT Basic Research Reference Number 6298: The analysis and synthesis of speaker characteristics

SOME ISSUES IN INVESTIGATING SPEECH STYLES

2. BACKGROUND

2.1. History

Much of the original work on style variation was conducted in the field of Social Psychology, and looked to correlate various aspects of a speaker's social background with their range of styles.

In 1972 [3] Labov introduced a psychological theory of speech style, which became a reference point for much of the subsequent work in this area. Speech was seen to inhabit a one dimensional range of styles, from most casual to most careful, with the position on this continuum determined by a combination of the amount of attention given to the task of speaking and the social status of the speaker. The implicit understanding was that speakers share an idealised speech norm and that more careful styles were better approximations of it. This can be seen to be related to Chomsky and Halle's [4] competence/performance distinction for language.

Along with this theory, for elicitation purposes, Labov suggested a range of tasks that could be equated, via the notion of attention, with different speech styles. These tasks spanned both scripted and un-scripted speech along a single continuum, suggesting that there is no difference, other than the speaker's attention, between these different registers. Labov also argued that emotive changes in speech were caused by the emotion's effect upon attention.

As might be expected, many aspects of this theory have been criticised, and Labov himself has expressed his doubts in it [5]. Milroy [1] gives a good summary of the evidence suggesting that scripted and un-scripted speech do not share the same idealised norm, and that a speaker's attention is not the sole cause of style variance. Bell introduced a competing "audience design" theory in 1984 [6] which suggested that variance was due to the way in which a speaker judged their audience. However, this seems to be equally weak in that it too allows only a single cause. The fact is that there are a large number of potentially causal factors to take into account with respect to style.

Thus speech style is dependent upon some internal speaker state, of which they may or may not be aware, which arises from contributory factors such as the speaker's attitudes, goals and the way in which they have interpreted their environment. The environment consists of an audience, a setting, and a task or field of discourse, and this will be interpreted in different ways depending upon certain aspects of the speaker's character — social factors, etc. The problem, then, is rather one of deciding how all these different and overlapping factors interact with each other and, when it comes to collecting data, what it is reasonable and possible to measure.

There are also some situations that have a set interpretation within a society, or language community, such as an interview, or rendition of a joke, and behaviour in these situations is more predictable and, therefore, more amenable to control. Different set situations can be expected to elicit speech in different registers.

If we look, then, at the types of speech that have been recorded, and the labels that researchers have used for them, we can distinguish at least two different aspects to which the term style has been applied. The first refers to the environmental conditions or task that was used to elicit the speech, such as scripted and un-scripted speech. And the second is concerned with the speaker's state — or perhaps with a listener's judgement of the speaker — as in casual and careful speech. This may, at least partially, relate to a confusion between style and register or a blurring of the boundaries between. That is, that perhaps read speech implies a different register? But then, the

text itself, could have been written in different ways and for different purposes. A play is obviously intended to be spoken, whereas letters usually are not. It seems as though we can make a large number of distinctions, and that they are not all mutually exclusive but depend upon our starting point: upon whether we are interested in behaviour, psychological states, or the social determinism of different situations.

There is need, then, for a new and comprehensive psychological model of speech and language variation that can accommodate these distinctions.

2.2. Speech Technology.
In recognition of the fact that a characterisation of different speech styles has potential use in both recognition and synthesis, speech technologists are now extending their work to consider the effects of style variation on speech. Researchers in this area have mainly restricted themselves to the differences between scripted and un-scripted speech, and Llisterri's paper [2] gives a comprehensive review of the style classification systems and the different recording conditions that have been used in recent studies.

2.3. Problems with Recording Speech.
The difficulty of recording anything approaching truly spontaneous speech, whilst exerting some control over what is said, is well recognised. And, understandably, eliciting the same words twice, for contrasting styles, adds to the problems. The term "un-scripted" has been adopted here to emphasise this.

An accepted method of eliciting un-scripted and unselfconscious speech is to set a speaker some task, where the goal and communication act are primary, and the actual words used incidental. However, to control what is actually said, without inducing a read style, the communication requirements of the task must be well constrained and the resulting messages reasonably predictable. An example of this can be found in work by Swerts [7], where speakers are asked to describe a sort of network diagram so that a listener can re-construct it from the recording alone. The diagrams consist of variously shaped and coloured objects connected on a network of lines. There is a limit, however, to the range of speech that any one task can tap into, and this may act to prevent the acceptance of standard tasks within the research community.

To compare contrasting styles the problem of repetition arises. The recorder must introduce another level of detail into the task in order to give the speaker a justifiable reason for repeating themselves. Work by Eskenazi [8] used recordings of telephone conversations, in which speakers were instructed to ring up a University department to make enquires about enrolling on a course. On the other end of the line a "wizard", acting as a departmental administrator, would get the speakers to repeat themselves, by immediately saying "Comment?" ("What?"), whenever they uttered one of the target phrases. The idea was to make the speaker think that their listener had not heard them properly and that they must repeat themselves more clearly.

More recent work at LIMSI has been directed towards the use of picture tasks for the collection of the French database of spontaneous speech to be used in the VOX project.

2.4. The Linguistic Variants.
The use of picture-based tasks to direct the informant towards particular words or phrases presupposes prior knowledge of the variants for the particular style change to be investigated. For English, vowel reductions, loss of voicing contrasts, loss of initial aspiration, final stop

devoicing/deletion and reduction of the /ing/ ending have all been found in casual speech relative to more carefully enunciated examples [9]. However, in selecting target contexts not only adjoining segments but also the presence of syllable, morpheme and higher level syntactic boundaries may play a part. From two [4] to five [10] different morpheme boundaries are believed to play a part in word-formation processes in English. This results in a vast number of potential target contexts for each possible variant.

3. DATA COLLECTION TECHNIQUES AT SHEFFIELD

Following the LIMSI example, the collection of English spontaneous speech for the VOX project, will involve a game of 'spot the difference', where the objects depicted and the differences between them are carefully selected to elicit target words and phrases, containing phonemic variables expected to vary with a change in style. Care has been taken in the planning of settings and tasks to try to ensure that informants will require the minimum of prompting.

3.1. The task and materials

At present, four pairs of pictures are used each with about seven differences that cover a range of the expected types of variance. The word "context" has been adopted here to refer to a phoneme sequence where some variability is expected. Each context may be a single segment or a series of segments, and may be morpheme- or syllable-internal or span a junction between words or subword units. Sixty target contexts were prepared for these four picture sets, but even this does not represent a balanced or complete range of the target contexts in all their possible syntactic positions. Further, English allows a number of different ways of naming the same thing, and whilst a lot of effort has been put in to developing pictorial materials most likely to elicit the target phrases, there is still an element of hit and miss for each speaker. However, four pairs of pictures does not present too difficult a task to speakers, and whilst we also hope to identify variations between the speech samples that have not been predicted, we do intend to extend the drawings to get a better linguistic coverage.

3.2. The session plan

Each speaker is recorded three times, to get examples of casual, careful and read speech. The speakers are informed that the objective is to study how people perform the 'spot the difference' game and encouraged to pay attention to detail. The first recording is presented as a practice run — intended to induce a casual style — for the second, "real", run in which the speakers are to be videoed. This provides a reasonably natural setting in which to request the speaker to repeat the task as closely as possible. The presence of a camera on the repeat run is intended to induce the shift from casual to careful speech. In reality however, speakers may be videoed during each of the phases and the recordings are expected to become a useful part of the database. During the interview to obtain details of the speaker's linguistic background, the casual phase of the recording is rapidly transcribed to provide the script for the read speech. For the main database collection it is hoped to used Speech Science interview rooms for the recordings in order to minimise the intrusiveness of the recording equipment. It is significant that, at the time of the data collection, the speaker understands that the recordings are to be analysed for their content, that is the picture task procedures, although, of course, no recordings will be retained or analysed without fully informed consent of the informant having been obtained.

3.3. The setting

During the game the 'experimenter' sits away from the speaker so that they are unable to see the

pictures. It is very easy to point out differences with your hands, head movements and almost no language at all, when both speaker and listener share the same view. This way however, the informant has only the verbal channel with which to get the message across.

One further, and quite important, detail is the context given to the speakers for these games. The speakers cannot be told the real reason for their performance as this will draw undue attention to the style shift and possibly affect it. For the French language recordings the speakers are told that the video is to be used to help teach the hard of hearing to lip read. They are also told, on account of this, to bear in mind how they speak during this videoed phase. However, for English speakers, it was feared that this might produce an over-articulation that equally affected the style. So for the preliminary recordings, the English speakers are given no reason for having to play the game beyond an interest in the game task itself. It is intended to include the 'lip-reading' instruction for some speakers in the preliminary phase in order to study the extent of this effect. This will be done by asking speakers who appear to exhibit noticeable style shifts to repeat the recording once again as we feel they may be suitable for lip-reading training videos. This issue is further discussed in the next section.

3.4. Establishing a base form for the description of variation

Once the speech is recorded, the data must be analysed and style differences characterised for use in the target speech synthesiser. For VOX purposes, then, characterisation must be in terms of parameters that the synthesiser can use. It is expected that this will be at the level of a phonemic segment description, with both casual and careful styles expressed in terms of variances from some "neutral" description. No claims are being made here about the psychological plausibility or significance of the "neutral" base. Further phonological analyses, including subsegmental descriptions, will be performed both at Sheffield and by making the databases available to the research community.

3.5. Speaker details

Relevant aspects of a speakers background information are taken once the first recordings have been obtained. Although certain factors will be ascertained for all speakers, the data is obtained through an interview rather than a questionnaire, as there are so many potentially relevant factors that may affect the way in which someone speaks. Thus, an interview style is adopted, in which the speaker is prompted to talk about themselves. Interesting themes are developed by the interviewer, and the potentially causal factors noted as part of the database. Once a large number of people have contributed to this data any recurring factor which seems to be important will be identified and any individuals whose range of answers is not complete will be recalled.

The work so far completed at Sheffield has been the development of task materials and of data collection methods which enable the collection of directed spontaneous speech for the analysis of contrastive styles. Other work in progress includes the development of computational tools for the style analysis phase.

4. DISCUSSION.

Among the issues that arise out of this work are the definitions of style, the choices of linguistic targets for analysis and the influence of the task and setting on the spontaneity of the speech elicited.

4.1. Definitions of Style
In section 2 we made the distinction between *task* and *state* based definitions of style, and used the terms *careful* and *casual* as examples of a *state* based definition. The aim of the VOX project is to study the differences between these two speech styles, but the recording technique that we propose relies on a contextual (task) change to elicit them. So how can we ensure that the recordings we obtain really are of careful or casual speech?

4.1.1. Observer Evaluation of Speech Style. Consider these three descriptions of speakers recorded for the French part of the project. They are anecdotal, based upon watching the video recordings rather than an analysis of the speech, but serve to illustrate the point.

(i) The first, M1, is an experienced teacher. He appeared very relaxed during both the casual and careful phase, but the pronunciation changed. In the careful phase, it was more 'exact' and his face was more animated, though the speed and pitch appeared to be the same. And whilst not looking up once during the casual phase, he looked up at the camera a number of times, when he knew it was on.

(ii) M2, an engineer unused to public speaking, appeared to be very self conscious during the careful phase. He looked up at the camera continually, and his speech lost its natural fluidity. The casual phase had been unremarkable.

(iii) The final speaker, M3, a student, seemed to be completely unaware of the camera. He did not look up once throughout and there was no noticeable change in his style.

Only one of these examples has produced the expected kind of shift from careful to casual speech, M1. The other two have either produced no real change at all, M3, or a wild and exaggerated change (into something that could better be described as "frightened" speech), M2. This leaves us with the question of whether to include the data from all three careful recordings in the same statistical class, or whether to discard M2 and M3 as they were not what we intended.

With Eskenazi's telephone recordings a jury, consisting of two phoneticians and the actual speaker, was used to decide if the careful speech was "making an attempt to be better understood" or not, and it seems that the evidence of a jury might make a useful contribution to our understanding of the perceptual judgement (or definition) of style. As the style characterisations are intended for use in a synthesiser then, a jury made up of the synthesiser's target audience could be used in VOX. There is no intention to discard data however. Instead, the response of the jury will become another parameter to be measured along with the speech and information on the speaker. If the jury is to try and judge the intention behind an illocutionary act however, we still have to decide how much of each phase to show them and exactly what questions to ask.

4.1.2. Theoretical definitions of Style. One of the original assumptions of Labov's theory was that all speech styles were successively better attempts at achieving an underlying ideal form. This is not the only possibility however. There may be representations of a number of ideal forms, for each different register, with both more casual and more careful versions of them all. The actual mappings from careful to casual may be similar for all the different registers. It is not even necessary to assume that there must be an "ideal" form. Instead the base representation could be of a casual form, along with a number of mappings onto more careful speech. Eskenazi [8] found a higher variance in some parts of the careful speech than in the casual, and used this as an argument to suggest that casual speech was fundamental. Another alternative would be to drop the

representational approach completely in favour of a connectionist type model. For the VOX project the determining factor will be the form of the representations and transformations required by the synthesis system. However, the data will also be examined for evidence to resolve some of these theoretical issues.

4.2. Task and Setting

Here we consider some of the factors which went into the design and selection of the picture tasks and the ways of presenting them to elicit speech without drawing attention to the speaker's manner of articulation.

4.2.1 The Linguistic Targets. In order to keep the picture tasks to a manageable level, target phrases were selected which incorporated multiple target contexts wherever possible. Thus, 'a red bucket with a green handle' contains at least six potential targets for comparison across speech styles. Further constraints on the selection of linguistic targets are imposed by the need for unambiguous pictorial representations to elicit the target phrases. However, all the data collected from each style will eventually be analysed to identify any segmental contrasts which occur in addition to those pre-selected.

4.2.2. Informants Perception of Task Purposes. First, we would like to consider the problem of what to tell the contributing speakers about the work. It helps to give meaning and plausibility to a task if participants are given a reason for what they are being asked to do. However, as a speaker's speech-awareness or self-consciousness affects the speech they produce, it is difficult to predict the effects of task instructions on the speech to be elicited. This further relates to the theory for style change which is assumed: reminding people of the fact that they are speaking is undoubtedly going to draw their attention to the process and, according to Labov [3], should be an effective way of making the speech style more careful. However, if a style shift is effected without the conscious attention of the speaker, this must have implications for the 'attention' theory? It will be interesting to compare examples of the speech produced when speakers have and have not been given the "lip reading" context. The results will also be compared with those of the French speakers to investigate the possible effects of cultural differences.

4.2.3. "Scripted Un-scripted" Speech. So far, we have not said much about the third phase of recording, where speakers read back a transcript of their own casual speech. As we pointed out earlier, a lot of research has looked into the differences between scripted and un-scripted speech. While it would be useful for us to have some read speech to be able to compare with the casual and careful styles, it is also important to enable us to evaluate these analyses with respect to previous research. However, we also recognise that there are many styles of read speech and that classifying these is, perhaps, as much of a problem as with un-scripted speech. Given this, it seems as though we could well obtain varying results, depending upon how the casual speech is transcribed and how we ask the speaker to read it back. Eskenazi [8] asked speakers to re-enact their telephone conversations to obtain the read speech. There are, however, no common situations in which someone would be expected to read back a transcript of their observations on a game of spot the difference. So the situation is already rather artificial, even before we decide how much alteration to allow in the process of making the speech more 'readable'. Alternative methods include treating it like a news broadcast, or, perhaps, inviting the informant to read the transcript like a policeman reading out a statement in court.

For our preliminary studies, because the task is rather artificial, we have tried to make the transcript as easy to read as possible. Each phrase is written on a new line, and the sentences are

amended by taking out the self-corrections and re-starts and changing some of the construction, as it would be for a news broadcaster. We need to develop methods which will enable us to identify the necessary interjections to minimise disruption of the natural rhythm of the speaker, but for the time being, we hope that this method will produce a uniform style that will suit most people.

5. CONCLUSION

Speaker style is not well defined, and there are many factors to consider when trying to analyse the ways in which speech varies. We have started recording data for the two styles of careful and casual speech. We are trying to make explicit as many as possible of the theoretical assumptions and decisions that our recording technique implies and, wherever we can, to treat any possibly relevant factor as another measurable variable. The drawback of this is that we might have to collect a lot of data before we can get many statistics. From the pilot recording studies it would seem that visual information is going to become a significant factor in attesting different styles.

In general there is a problem with recording style data, in that we must use either a task (environment) dependent or listener dependent definition for what we are assuming to be a speaker-internal state.

6. ACKNOWLEDGEMENTS

To the members of the VOX consortium and in particular to our French partners at LIMSI for their contribution and their inspiration for the work we have implemented here at Sheffield: Maxine Eskenazi and Vincent Pean.

7. REFERENCES

[1] L MILROY, 'Observing and Analysing Natural Language'. Oxford, Basil Blackwell, (1987).
[2] J LLISTERRI, 'Speaking styles in speech research', ELSNET/ESCA/SALT workshop on 'Integrating Speech and Natural Language', Dublin, Ireland, (1992).
[3] W LABOV, 'Sociolinguistic Patterns'. Philadelphia: Pennsylvania University Press, (1972).
[4] N CHOMSKY & M HALLE, "Sound Pattern of English", Harper & Row, (1968).
[5] W LABOV, 'Field methods used by the project on linguistic change and variation', Sociolinguistic Working Paper 81. Austin, Texas: South Western Education Development Laboratory (1981).
[6] A BELL, 'Language style as audience design', Language in Society 13(2), pp 145-204, (1984).
[7] M SWERTS, 'An experimental approach to the study of prosody in spontaneous speech', ESCA workshop on 'Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication'. Barcelona, Catalonia, Spain, (1991).
[8] M ESKENAZI, 'Changing speech styles: strategies in read speech and careful and casual spontaneous speech'. Int. Conf. on Speech and Language Processing, Banff, (1992).
[9] A D ZWICKY, 'Styles', in: "Style and Variables in English", eds.Shopen. T., Williams, J.M., Winthrop Publishers Inc., Cambridge, Massachusetts, (1981).
[10] M HALLE & K P MOHANAN, 'Segmental Phonology of Modern English', Linguistic Inquiry 16(1), (1985).