



HUMAN FACTORS ASPECTS OF TEMPLATE TRAINING

K.Hapeshi^o, D.M.Jones^o, C.Frankish*.

ABSTRACT

The success of speaker-dependent speech recognition systems will largely depend upon the consistency between utterances made during reference template training and those made during system application. Inconsistencies can result from differences in the environmental conditions, the behaviour of the user or because of the discrepancy between task demands made during the application compared to template training. A careful scheme of user induction can help, but the aim for designers must be to prepare system procedures that help utterance consistency. Suggestions are made for doing this by improving methods of harvesting utterances on which reference templates will be based.

METHODS OF TEMPLATE TRAINING

Most automatic speech recognition (ASR) devices are speaker-dependent, and hence require user to verbalize each vocabulary item a number of times. In this way users 'train' reference templates against which future utterances are compared. The number of repetitions required for different devices vary quite considerably. For example for the Votan VPC 2000 two templates are sufficient, while for the Kurzweil KVS five to ten repetitions (referred to as 'tokens') are suggested. The number of repetitions per vocabulary item is very important since adding more tokens is considered to be the surest way to improve recognition accuracy (Wilpon, Rabiner and Bergh, 1982; Spine, Williges, and Maynard, 1984).

Template training methods may also differ in the way users are expected to produce sample utterances since reference templates could be based on isolated or connected speech. In general, templates extracted from connected speech are more robust; however, a major problem with current connected speech recognizers is the time it takes to train templates. When comparing an isolated with a connected speech recognizer, Christ (1986) found that training times increased four-fold, because each item had to be trained in a random context as well as individually for the connected recognizer. Christ reported that during the 40 minutes for connected training subjects became bored and anxious for the session to finish. Thus the method of template training will significantly affect the duration of the session, and it is likely that long and boring template training sessions will have a detrimental effect on the quality of utterances made by users.

SOURCES OF RECOGNITION ERRORS

The method used for template elicitation is important because the most common source of recognition errors is inconsistency between verbalizing items during template training and verbalizing them during use of the system. In general three reasons for inconsistencies can be identified: Firstly, inconsistency can result from the discrepancy between

^o Department of Applied Psychology, University of Wales Institute of Science & Technology, Cardiff CF3 7UX.

* Department of Psychology, University of Bristol

system and application characteristics. Secondly, there are problems resulting from the cognitive or physical limitations of the user. Thirdly, there are problems resulting from the attitude of the user towards the system. These will be considered in turn.

Application and system characteristics

Hardware inconsistencies, such as using the wrong microphone or input gain, are relatively easy to avoid; however, inconsistencies caused by the environmental conditions in which the ASR system is used are more difficult to control. With many systems, template training is carried out under different circumstances to those in which the system will be used. Generally, training may be carried out under relatively ideal (quiet, private, and comfortable) conditions, while the system is to be used in noisy, busy, and uncomfortable or stressful conditions (Green, Payne, Morrison and Shaw, 1983). Also template training is often an artificial and repetitive process, with speakers required to repeat items in list form; during the ASR application the system and user may be involved in a relatively more natural and continuous dialogue (Barry and Williamson, 1986).

Physical limitations of the user

For some applications the user is expected to make use of the speech recognition device throughout the working day, or at least for long periods. For example, there are currently a number of attempts to develop large vocabulary automatic speech recognizers for word-processors or devices for dictating letters (Baker, 1986; Cole, 1986; Kurzweil, 1986; Meisel, 1986). However, little consideration has been given to the problems of fatigue using these 'talkwriters' for extended sessions (Newell, 1984). It is certain that users would suffer from vocal fatigue, but it is also possible that there will be general psychological fatigue, and in both cases these could affect the voice of the user. Prolonged speaking, resulting in voice fatigue, is also likely to be a problem for applications in which the speech device may be used only occasionally, but the user is speaking to people for long periods. For example, in a study at the New York Stock Exchange, Zarembo (1986) reported that the voices of clerks working on the floor became strained with use by late afternoon, seriously affecting the recognition performance of the speech input system. One solution is to store a complete set of templates for the end of the day to avoid retraining (Connolly, 1979; McCauley, 1984; Zarembo, 1986). Alternatively, if the ASR device uses an averaging technique, users can input one or two samples of critical items each time the system is used to update the templates (Lynchard, 1981).

Perhaps the most important class of user limitations that must be considered when designing speech recognition systems, are those concerning the human information processing system. For example, it may be difficult for users to remember the whole of the vocabulary necessary to use the system. The need to memorize items can to some extent be eliminated by allowing users to use terms they are familiar with such as "rubout" instead of "delete". In a study reported by Lind (1986), airport baggage handlers were encouraged to use the destination names they were most familiar with (for example, "Frisco", "LA", "Kennedy"). This gave users more confidence and allowed for a more readily accepted system, while new users could learn how to use the system to an adequate degree in just 15 minutes.

Allowing users to adopt their own vocabulary may result in a more easily learned and 'user friendly' system. However, this may cause problems in some applications when recognition errors occur, especially if

it is necessary to revert to another mode of input such as the keyboard. In these cases users may be required to remember an alphanumeric code that they rarely use, or had not used for many weeks or months, and so may have difficulty avoiding errors. Therefore, 'user-friendly' systems, which allow users to select their vocabularies, might create more problems than they solve. The solution may be to design a system of prompts or easily remembered backup codes to help the user when reversion to another mode is necessary.

Users' attitudes towards an ASR systems.

Problems in speech recognition can occur as a result of inconsistencies in the users' attitudes towards the system during training and when actually using the system. For example, Nelson (1986) reported that in the initial stages of using a speech input system for inspection, errors were due to giggles, coughs, and general self-consciousness on the part of the users. Nye reported noticeable improvements in performance after a week (when users became more experienced with the system) voice input was as fast as manual input, and these continued for a further six weeks when voice input was on average 30% faster than manual input.

The time it takes for users to adopt a more appropriate and consistent attitude towards a system is an important factor which will determine the success of the application. Certainly in some applications, it is important that the user is aware of how the system works, but that they should not try too hard when using a system. Nye (1982) pointed out that one of the most common errors which users make is to overpronounce utterances in order to 'help' the speech recognition device. This would not be a problem if the user was consistent, but when the user is more familiar with the system, he or she may speak more normally, thus causing recognition errors. Also, an ASR system can work very well once users have become familiar with it and are pronouncing words normally and carefully, but after an initial excitement, slurring or pronunciation errors reappear, which can mean that all or part of the vocabulary must be retrained.

IMPROVING GENERAL USER CONSISTENCY

Studies have shown that user consistency between template training and application can be improved by instructions designed to help users maintain a 'good' voice pattern while they use an ASR system (Wilpon & Roberts, 1986). Manufacturer's recommendations include suggestions to lengthen the utterance by speaking more slowly on short items such as single syllable words or digits (KVS Manual, 1985; VPC 2000 Users Guide, 1985). In this way the algorithm is more likely to be able to extract features in the speech utterance. False recognition with longer, multiple syllable utterances is more likely to be the result of inappropriate pauses within the utterance, which may falsely signal the end of an utterance to the recognition algorithm. Therefore, the user is told not to pause within a phrase item, but advised to speak "crisply", "clearly", "quickly" (on multi-syllable utterances), "with rhythm", and "with normal loudness".

Another more obvious measure is to avoid rushing between items, causing them to be merged and result in false recognition, or false rejection. For example, saying the phrase "there are four" very quickly could trigger false recognition of the item "therefore". A tendency to merge items can be overcome by extracting templates from samples of continuous speech during training if the ASR device allows this. However, for connected speech recognition to work well speakers must stress each

word or word-equivalent with no significant co-articulation between adjacent words (Meisel, 1986). An alternative method for improving recognizer performance is to place the ASR device in connected speech recognition mode during system application, even if template training was carried out in isolated recognition mode. Paradoxically, this results in better recognition accuracy since the software makes no assumptions with regard to the beginning and end of the utterance, instead it tries to maximise the match with stored templates (Tincello, 1987).

Improving consistency during ASR application

Despite general user training even the most experienced users suffer from a degree of change or 'drift' in their voice over time (Green, Payne, Morrison & Shaw, 1983). If this becomes serious then templates may need to be completely retrained. However, it may be possible to reduce voice drift by providing adequate feedback to users to indicate how closely utterances match the reference templates. Even relatively simple forms of immediate feedback have been shown to improve system performance in terms of recognition accuracy (Poock, Martin & Roland, 1983; Schurick, Williges & Maynard, 1986). Although feedback can result in better overall performance, it is not clear if it can improve speaker consistency per se. Wilpon and Roberts (1986) reported that a simple barometer-like display indicating distance scores between utterances and the best match template, did not improve speaker consistency. However, Wilpon and Roberts only provided feedback if a pre-set difference threshold was not reached. Therefore feedback was essentially given as a form of negative reinforcer, which is known to be a poor way of encouraging a desired behaviour (Skinner, 1953). If feedback was given after each utterance, then it may be more effective as a positive reinforcer of utterance consistency.

Improving template training

Some recognition problems resulting from inconsistencies between template training and system use can be avoided if certain precautions are taken at the time of training. In the Votan VPC 2000 User Guide (1985 version) users are advised to train at midday when the voice has "warmed up", but is not yet "tired". However, as Waterworth (1984) pointed out, a major problem for current speech input systems is that the man-machine dialogue environment during template training is not the same as that during use in the speech recognition application. In many applications within-speaker variability is a serious problem, particularly when speakers must perform concurrent tasks (Cooper, 1987). To overcome this template training can be made to simulate the conditions that prevail during system application so that utterances are similar. For example, it has been suggested that when within-speaker variability is inevitable and changing environmental conditions cannot be avoided, then variability can be deliberately 'built-into' the reference templates (Poock, 1980; McCauley, 1984). Thus users can be encouraged to vary their utterances during template training, or to provide templates at different times of the day, and under different conditions.

The main aim for human factors researchers must be to devise training methods that ensure variability is accounted for placing the onus for consistency on the system and not on the user. Also, with advances in software algorithms, it may be possible to make future systems insensitive to variability, or at least automatically adapt to gradual voice changes over time (Green et al, 1983). Until then, the success of ASR applications relies on the ability of the user to provide representative templates and remain consistent with these. With current systems users must learn how to be consistent through their experience with using the speech recognition

devices. It might be possible to disguise training to simulate the application, which may help to make the session more interesting and meaningful (Green et al, 1983). However, if subjects are unaware that the initial session is only template training, they might believe they can do productive work and become frustrated or disillusioned (Waterworth, 1984).

CONCLUSIONS

A number of improvements can be incorporated into template training procedures to increase consistency with the conditions that prevail during system application. Firstly, template training should be carried out in the same environmental conditions using the same utterance sequences required by the application. Secondly, users should be subjected to the same work-load stresses during the template training session as they would be exposed to during application, including those that produce fatigue. Thirdly, users should be provided with consistent feedback with respect to voice consistency, both during template training and application of the system.

Research on refining these improvements and incorporating them into template training procedures are being carried out by on an Alvey funded project described elsewhere (Jones & Hapeshi, 1987; Jones, Hapeshi & Frankish, 1987).

REFERENCES

- T P Barry & D T Williamson, IEEE, 799 (1986)
J M Baker, Speech Tech '84, 22 (1984)
K A Christ, Speech Tech '86, 262 (1986)
A Cole, Speech Tech '86, 149 (1986).
D W Connolly, Nat Av Fac Ex Cen Tech Rep FAA-NA-79-20, (1979)
M Cooper, Speech Tech, 3(4), 82 (1987)
T R G Green, S J Payne, D L Morrison & Shaw, Beh & Inf Tech, 2, 23 (1983)
D M Jones & K Hapeshi, Int Speech Tech '87, May (1987)
D M Jones, K Hapeshi & C Frankish, HCI '87, in press (1987)
R Kurzweil, Speech Tech '86, 184 (1986)
A J Lind, Speech Tech '86 (1986)
R Lynchard, Cited by McCauley, 1984, (1981)
M E McCauley, Hum Factors Rev, 131 (1984)
W S Miesel, Speech Tech '86, 189 (1986).
A Monk (ed), Fund of Hum-Comp Inter, (Academic Press, 1984)
D L Nelson, Speech Tech '86, 62 (1986).
A F Newell, INTERACT '84, 174 (1984).
J M A Nye, Speech Tech, April, 50 (1982)
G K Poock, NPS Tech Rep NPS-55-80-016 (1980)
G K Poock, B J Martin E F & Roland, NPS Tech Rep NPS-55-83-003 (1983)
J M Schurick, B H Williges, & J F Maynard, Ergonomics, 28, p1543 (1985)
B F Skinner, Science & Human Behaviour (Macmillan, NY, 1953).
T M Spine, B H Williges & J F Maynard, I J Man-Mach Stud, 21, 191 (1984)
D Tincello, personal communication with authors (1987)
J A Waterworth, In A Monk (1984), p221-36
J G Wilpon, L R Rabiner & A F Bergh, J Acous Soc Am, 72(2), 390 (1982)
J G Wilpon & L A Roberts, IEE (1986)
C A Zarembo, Speech Tech '86, 69 (1986)