Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 42 (2004) 289–312
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Modelling of confusions in aircraft call-signs

Stephen Cox [a,*], Lluis Vinagre [b]

[a] School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK
[b] National Air Traffic Services Ltd., London Terminal Control Centre, Porters Way, West Drayton UB7 9AX, UK

Received 6 January 2003; received in revised form 8 July 2003; accepted 8 September 2003

Abstract

Air-traffic has grown rapidly in the last twenty years and concern has been mounting about the safety implications of mis-recognition of call-signs by both pilots and air-traffic controllers. This paper presents the results of a preliminary study into perceptual (i.e. non-cognitive) confusions in two closed vocabularies of the type used as aircraft call-signs. Conventional methods of subjective and objective testing were found to be unsuitable for our aim of predicting potential confusions within a vocabulary. Hence a method for modelling confusion probability in a closed vocabulary at a certain signal-to-noise ratio has been developed. The method is based on the use of a phoneme confusion matrix and a technique for comparing phoneme strings. The method is presented and results are given. These suggest that the behaviour of the model is plausible, and a comparison of its predictions with a set of real confusions showed a correct prediction of position of confusion in three-word phrases. The predictions of the model need to be verified by subjective testing before it can be deployed in a system that designs low-confusability call-signs, which is the ultimate goal of the research.

1. Introduction

There has recently been concern in organisations concerned with air-traffic control about safety incidents resulting from the actual or potential confusion of airline call-signs. A recent Aeronautical Information Circular (Services, 1996) stated that "Whilst [the CAA Mandatory Occurrence Reporting Scheme] has established that there are definite safety implications resulting from call-sign confusion, a dedicated study has not been conducted." This study is a preliminary investigation into some aspects of this problem. A typical format for an aircraft call-sign is three letters, which designate the aircraft operator, followed by two to four digits (or a combination of digits and alphanumeric characters) which are specific to the flight. Examples of typical call-signs are BAW 602, DAL 41 etc. An aircraft controller may be directing as many as twelve aircraft at any one time and communicating with their pilots on a single radio-telephone link: hence there is potential for confusion if the call-signs are similar. The CAA maintains a database of actual and potential call-sign confusions. Some examples of the kinds of confusions that this contains are given below (N.B. the confusions given below involved only the digits section of the call-sign and the initial three letters are not shown).

* Corresponding author.
E-mail address: sjc@sys.uea.ac.uk (S. Cox).

0167-6393/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2003.09.006

Words substituted:	TWO-OH-TWO	→	TWO-OH-THREE
	SEVEN-OH-NINE	→	SEVEN-OH-EIGHT
Words transposed:	TWO-SEVEN-EIGHT-NINE	→	TWO-EIGHT-NINE-SEVEN
	ONE-OH-THREE	→	THREE-OH-ONE
Words inserted:	SEVEN-OH-ONE	→	SEVEN-OH-OH-ONE
Words deleted:	ONE-THREE-SEVEN	→	THREE-SEVEN

The cause of these confusions can be divided into two main effects that are to do with early and late processing in the brain. The first effect, which is concerned with early processing, is a perceptual one: the phrase spoken was mis-recognised because it was poorly articulated, or because there was noise in the communication channel when it was spoken, or because the listener's hearing is poor (Vandeelen and Blom, 1990), or because his/her English is poor etc. The second effect, which is to do with later processing, is a cognitive one that is due to short-term memory. Classically, information is lost from short-term memory by the processes of displacement (existing information is replaced by newly received information when the storage capacity is full), decay (information held in a "register" needs to be maintained or it decays over time) and interference (other information in storage distorts the original information) (Baddeley, 1990). To these processes can be added another important effect which is conscious or unconscious "filtering" of the message because of prior expectations e.g. "They always ask me to descend at this time", "He must have meant TWO-SIX-ZERO because there's no aircraft called TWO-EIGHT-ZERO" etc. It is likely that the perceptual and cognitive effects will interact. If the signal quality is poor for any of the reasons mentioned above, the listener may exhibit weaker memory of the message, or may be more prone to psycho-linguistic errors, or may be more liable to fall back unconsciously on what he or she expects to hear.

1.1. Scope of this study

This study has focussed on the first effect, the perceptual one. This is likely to be a simpler effect to study and to quantify than the cognitive effect. Our approach has been to define two artificial but plausible call-sign vocabularies that consist of words commonly used in this task, and to use a model of speech perception to simulate the confusions within these two sets. We prefer to begin by using a model-based approach rather than the more direct approach of testing listeners on real speech signals. The reason for this is that although the model may not be as accurate in predicting confusions as direct testing, it allows us to test comprehensively a very large vocabulary at a number of different signal-to-noise ratios (SNRs), which would be very expensive and time-consuming to do using listeners. The aim of this simulation is to identify the main effects and the kind of problems that might be expected from the call-sign vocabularies. The information gained from this study will then enable us to devise a subjective test that is much smaller in scope and that focuses on these effects. The results from the simulation will also form the basis of the first stage of the ultimate goal of the work, which is to provide a tool for air-traffic controllers to design (dynamically) low-confusability call signs so as to minimise perceptual error amongst the group of pilots under their control at any time.

Two sets of phrases were used:

1. Groups of three digits e.g. FIVE-SEVEN-ONE, TWO-FOUR-EIGHT etc. (the *digit-triple* (DT) set).
2. A single letter from the airline alphabet followed by a single digit e.g. ALPHA-FOUR, KILO-ZERO (the *alphabet-digit* (AD) set).

The international aviation alphabet (ALPHA, BRAVO, . . . ,ZULU) was mandated by the International Civil Aviation Organisation (ICAO) when English was adopted as the global language for aviation. The words were apparently not chosen on the basis of their phonetic properties: rather, they were chosen because they were familiar,

and easily pronounceable by non-English speakers. More information on the alphabet is given in (ICAO, 1990).

Ideally, a full call-sign of three airline alphabet letters and three digits would have been modelled, but this would have made the study infeasible, requiring the comparison of approximately 24 000 000 different call-signs. It is clear that the digits are likely to be potentially the most confusing part of a call-sign, and so the first vocabulary was chosen to test confusability of digit triples. The second vocabulary was chosen to evaluate the effectiveness of the longer and phonetically richer call-signs in combating confusion. To avoid confusion, a particular set of phrases such as the DT set and the AD set is referred to as a *phrase-set*, and the set of words that comprise all the phrase-sets as the *vocabulary*.

Two factors that affect intelligibility of speech were studied:

1. Broadband noise added to the speech signal.
2. The effect on confusion performance of co-articulated speech i.e. speech which is pronounced spontaneously, rather than speech according to the "canonical" pronunciations given in a dictionary.

Again, these are not comprehensive: other effects such as that of channel bandwidth, of different kinds of interfering noise in the channel, of non-native accents etc. could also have been included, but these must be left to a later study.

2. Use of established techniques

The advantages and disadvantages of the two ways of measuring the performance of a speech communication system are well-known. Subjective measurement requires panels of listeners to give their responses to stimulus words that are spoken over the system. Both stimulus and response words are usually selected from a closed list of words (such as the modified rhyme test, the diagnostic rhyme test or phonetically balanced words), although open responses are sometimes also used. Such testing gives results that are reliable, but it is very expensive. Objective measurement uses techniques such as the articulation index (French and Steinberg, 1947), the speech transmission index (STI) (Steeneken and Houtgast, 1980, 2002), the rapid STI (RASTI) (Steeneken and Houtgast, 1985) or the speech intelligibility index (Mendel et al., 1998) which measure the response of the system to a special test signal and attempt to predict the intelligibility from analysis of this response. Objective measurement is much cheaper, but the intelligibility predictions it produces are less reliable, and the need for calibration and the effect of possible inaccuracies may entail costs. In addition, these techniques measure only the intelligibility of a system, which is usually quoted as the average percentage of words that will be correctly understood by a user of the system (although some measures (e.g. STI) output an intermediate index which then can be related to different indices of intelligibility, e.g. to sentence scores, PB word scores, CVC word scores, etc.).

In this study, our aim was to make a closer examination of confusion effects that occur in a certain phrase-set rather than merely estimating the intelligibility. In particular, we wished to estimate the probability of each word or phrase within a closed phrase-set being mis-recognised, so that potentially troublesome phrase-set items could be identified. This information is represented as a *confusion matrix*, in which element $C(i,j)$ of the matrix gives the probability of the response being item j in the phrase-set when the stimulus was item i. The confusion matrix of the system is clearly a much more informative measurement about the system than the intelligibility—the intelligibility can be calculated from the confusion matrix but not vice-versa. The objective methods of testing mentioned above are not suitable for this study as they are not capable of producing a confusion matrix. Subjective testing is also highly problematical for this study given the large number of possible call-signs in a phrase-set. The size of the two vocabularies is respectively 1000 and 260 phrases, and these should ideally be tested at a number of different SNRs using at least ten listeners. This is impractical unless a highly reduced subset of the phrase-set is used, which then raises the question of the validity of the results.

2.1. Predictive modelling of confusions

In 1977, Moore considered the problem of how to assess the performance of speech recognisers tested on different vocabularies (Moore, 1977). He attempted to create a universal metric that was independent of the vocabulary used to test a recogniser by benchmarking recogniser performance against human performance. His idea was to measure the performance of a recogniser in terms of its Human Equivalent Noise Ratio (HENR), which is the signal-to-noise ratio (SNR) of the speech material that would be needed to degrade a human's recognition performance to be the same as that of the machine. Hence a recogniser that had a high HENR rating would be a good one (equivalent to the performance of a human listening at high SNR)) and performance would drop as the HENR dropped. Another useful metric that he introduced in this work was stress, which is a measure of how different the machine's confusion matrix on a certain vocabulary is from that of a human. In order to be able to measure HENR and stress, Moore required a way of predicting the confusion matrix for a human on a given vocabulary at a certain SNR.

At that time, there were a few studies of confusions of consonants and confusions of vowels available. The most useful was a study by Miller and Nicely (1955) of consonant confusions at different bandwidths and different SNRs. There were also papers by Peterson and Barney (1952) and Pickett (1957) on vowel confusions, although these were less complete in terms of their SNR coverage than the consonant confusion work of Miller and Nicely. However, even this work was insufficient to build a full model of consonant confusion as it covered only 16 of the consonants, whereas something like 24 are required for full coverage of English words. Hence Moore used data from some other studies (Singh et al., 1972; Wang and Bilger, 1973) and ingeniously integrated it with Miller and Nicely's data using a multi-dimensional scaling technique. The result was a model that could predict a confusion matrix for consonants and a confusion matrix for vowels at any required SNR. Having constructed these two matrices, he used them with another model to predict the confusion matrix for an isolated word vocabulary.

Moore's work was extended to predicting speech recognition accuracy in a study by Simons (Simons, 1995). Simons was interested in the problem of predicting the recognition accuracy of a speech recogniser on a certain vocabulary, spoken in isolated word fashion. Rather than building confusion matrices from data gathered from experiments using listeners, she used the phoneme confusion matrix generated by the recogniser together with the technique of dynamic programming to produce a confusion matrix for the vocabulary words. This was done at only a single SNR. The results were very encouraging: her final system achieved a correlation of 0.95 between predicted and measured accuracy on a given vocabulary.

2.2. Selection of a testing technique

The conventional methods of subjective and objective measurement are problematical for this study for the reasons given in Section 2. Moore's technique is attractive because it enables prediction of confusion performance rather than simply intelligibility. Furthermore, Moore verified it using a panel of 11 subjects on a vocabulary of 40 words, and found a "diagonal rank correlation" of 0.73, which was judged to represent a good prediction by the model. Although it has not been verified for human performance, the work by Simons shows that a similar technique gave excellent results in prediction of accuracy for an automatic speech recogniser. It was therefore decided to use a synthesis of the ideas of Moore and Simons to predict confusion performance. However, predictions from the model will have to be verified by listening tests to establish the validity of the model. The details of the techniques used are given in Section 3.1.

3. Modelling technique

3.1. Overview of technique

The probability of confusion of a certain phrase-set is estimated by producing a *phrase-set*

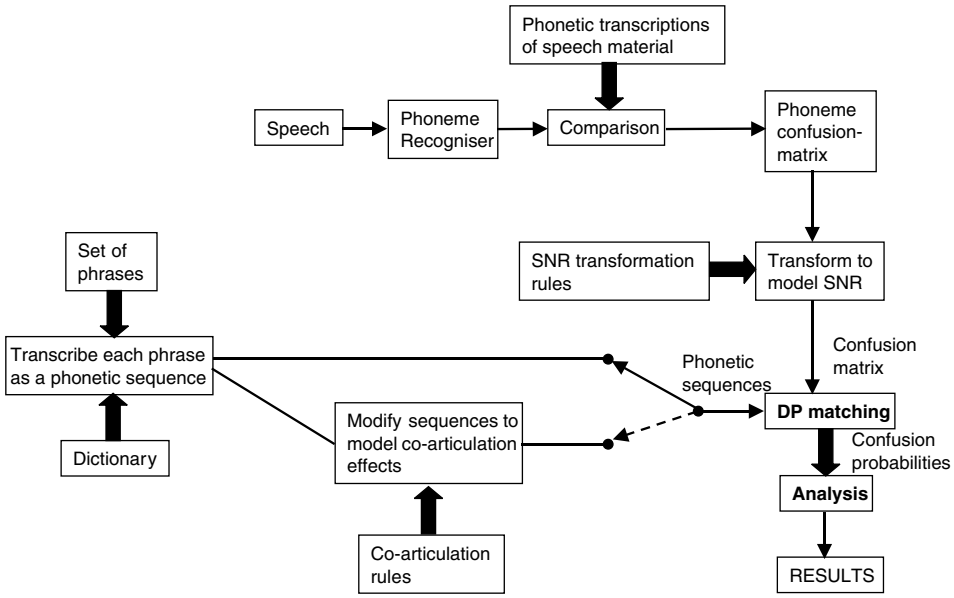

Fig. 1. Overview of the complete process used to generate a confusion matrix.

confusion matrix, C. C is an $N \times N$ matrix (where N is the number of phrases in the phrase-set) which records the probability $C(i,j)$ of the "response" phrase being R_j given that the "stimulus" phrase S_i is input. The essential steps in producing C for a certain phrase-set at a certain SNR are shown graphically in Fig. 1.

1. Use a dictionary to transcribe each phrase in the phrase-set into a phonetic sequence.
2. If required, manually edit these sequences to reflect more realistic pronunciations in rapid speech using the co-articulation rules described in Section 3.6.
3. Use a phoneme recogniser to recognise a standard speech database.
4. Use the phonetic transcriptions of the items in the database to obtain a phoneme confusion matrix (Section 3.2).
5. Model the effect of a certain SNR on the confusion matrix using the processes described in Section 3.5.
6. Use dynamic programming (DP) together with the confusion matrix to produce a matrix of confusion probabilities for the phrase-set (Section 3.3).
7. Normalise and analyse the probabilities (Section 3.4).

The studies by Moore and Simons differed in steps four, five and six. In steps four and five, Moore used "human" confusion matrices for vowels and for consonants predicted by his model at different SNRs, whereas Simons used a single confusion matrix from a speech recogniser. In step six, Moore used a deterministic approach based on the rules of English syllabic structure for matching phones between corresponding syllables, and a rather *ad hoc* approach for lining up syllables, whereas Simons used dynamic programming.

3.2. Choice of confusion matrix

In this study, it was decided to use a confusion matrix produced by an automatic speech recogniser rather than confusion matrices derived from experiments on listeners. There were several reasons for this:

1. Although the human confusion data for consonants from Miller and Nicely's study is reasonably comprehensive (16 consonants at 6

different SNRs tested on 10 listeners), it does not cover all the consonants and, in addition, the vowel data is very sparse and only available at a single low SNR. Moore overcame these problems ingeniously in his model, but doubts must remain about the assumptions he was forced to make in doing so.
2. There is no human data for consonant/vowel confusions available.
3. A confusion matrix generated by a recogniser is derived from speech from an order of magnitude larger number of speakers than the number of speakers and listeners used to generate the human confusion matrices, and so may be more representative of real confusions.

The obvious objection to the proposal to use a confusion matrix generated by an automatic recogniser is that the kind of mis-recognitions made by an automatic recogniser are different from those made by a human. This point was checked carefully in this study, and our conclusions are that this was not the case: the automatic recogniser has a pattern of confusions similar to humans. Section 5.2 takes up this question in much more detail.

3.3. Comparison of phoneme sequences

A second problem was how to compare two phoneme strings using a certain phoneme confusion matrix. Comparison of words that have the same number of phonemes is straightforward. Consider, for instance, computing the probability that the word BIT (/b ih t/) is recognised given that POD (/p oh d/) is spoken i.e. it is required to calculate Pr(R = b ih t|S = p oh d). [1] Using the notation Pr(R = a|S = b) to mean "the probability that the response is a given that the stimulus is b", this probability can be estimated by assuming that the three events Pr(R = b|S = p), Pr(R = ih|S = oh) and Pr(R = t|S = d) are independent, so that Pr(R = b ih t|S = p oh d) = Pr(R = b|S = p) × Pr(R = ih|S = oh) × Pr(R = t|S = d), and these three probabilities can be looked up in the confusion matrix. The independence assumption has been verified for CVC nonsense syllables (Fletcher, 1953), but is unlikely to hold for words. However, it would be impractical to estimate joint or conditional probabilities for groups of phonemes.

When there are unequal numbers of phonemes in the two words, Moore considered that the syllable was the important unit and developed a set of rules for matching syllables. In his formulation, when a monosyllabic word such as THREE is matched to a disyllabic word such as ZERO, the "extra" syllable in ZERO is matched to the repeated first syllable i.e. Pr(R = THREE|S = ZERO) = Pr(R = THREE|S = ZE) × Pr(R = THREE|S = RO). Moore stated that these rules had no data to substantiate them and this rule seems incorrect, as it forces matching of events that occur at different times in the utterance.

Since Moore's paper was published, dynamic programming (DP) has been extensively used in speech processing to align both speech segments and symbol sequences of different lengths. DP will produce the optimal alignment of two sequences according to a specified criterion, such as minimum overall Euclidean distance. Each entry in a phoneme confusion matrix can be regarded as the probability of a "response" phoneme given a "stimulus" phoneme. Hence if one of the phrase-set phoneme strings is regarded as the stimulus and the other as the response, the criterion "maximum response probability" will find the DP alignment of the two sequences that produces the highest response probability.

Consider the problem of matching THREE (/th r iy/) with ZERO (/z ia r ow/). One possibility is to allow the introduction of a null phoneme (#), as follows:

th	r	iy	#
z	ia	r	ow

On the assumption that phoneme confusions are independent, an alignment that has a higher overall probability according to the confusion

[1] In this paper, the computer-readable ARPAbet symbols are used to provide a broad phonetic transcription of the words in the lexicon. Appendix A gives an equivalence between ARPABET and IPA symbols.

matrix is as follows:

th	#	r	iy
z	ia	r	ow

There are some problems (identified by Simons) with the idea of using a null phoneme for matching and we prefer not to do this. Hence it is required to repeat either the th or r phoneme in THREE, and the DP algorithm chooses which to repeat on the basis of "maximum response probability".

The merit of using DP to compute the similarity between two phoneme sequences is that *ad hoc* assumptions about which the important phonemes are, or how phonemes may or may not match are not used. Rather, the matching is done by the principle of maximising response probability. This may produce some alignments that are implausible from a phonetic point: however, we argue that this may not be completely undesirable. An automatic procedure of this sort cannot be expected to produce results that are as accurate as testing on humans. If they are inaccurate, it would be better for them to err by reporting a higher rather than a lower mis-classification probability than was actually the case. Because the DP procedure seeks to maximise the probability that a stimulus phoneme string S is mis-recognised as R, it may have the effect of boosting the mis-classification probability.

3.4. Normalising the confusion matrices

Consider a simplified situation in which there are only two "phonemes" in the language, X and Y, and only two "words" in the phrase-set, XY and YX. Suppose the phoneme confusion matrix is as shown in Table 1.

The phrase-set confusion matrix is then as shown in Table 2.

Notice that both rows of the phrase-set confusion matrix sum to 0.62, not 1.0. This is because there are two other possible "words", namely /XX/ and /YY/, which are missing from the phrase-set. If the probabilities $\Pr(R = XX|S = XY) = 0.24$ and $\Pr(R = YY|S = XY) = 0.14$ are added to row one, the sum is 1.0 as expected. These missing responses can be accounted for by assuming that the proportion of the total probability in row i held by element (i, j) of the confusion matrix is what matters, and to normalise each element in row i by dividing by the sum of the elements in row i. Hence the phrase-set confusion matrix becomes as shown in Table 3.

Table 1
An example phoneme confusion matrix

Input	Recognised	
	X	Y
X	0.8	0.2
Y	0.3	0.7

Table 2
The phrase-set confusion matrix

Input	Recognised	
	XY	YX
XY	$0.8\times0.7 = 0.56$	$0.2\times0.3 = 0.06$
YX	$0.3\times0.2 = 0.06$	$0.8\times0.7 = 0.56$

Table 3
Normalised phrase-set confusion matrix

Input	Recognised	
	XY	YX
XY	0.9	0.1
YX	0.1	0.9

Prior to any normalisation, a phrase-set consisting of phrases made up of long phoneme sequences will produce a row of response probabilities that are lower in value than probabilities produced by another phrase-set that is similar in every other way except that it consists of phrases that are made up of shorter phoneme sequences. However, what is important is the *relative* response probabilities for a given stimulus phrase, and normalisation along a row will tend to reduce the difference between vocabularies of different lengths. Normalisation was applied to all the phrase-set confusion matrices computed in this study.

3.5. The effect of noise

In this study, the effect on the confusion matrix of decreasing the SNR was modelled in two ways. The first and simpler way was to assume that the effect of additional noise is to re-distribute probability "mass" from a diagonal element of the matrix to elements along the corresponding row. It seems reasonable to use a model in which the relative probabilities of confusion amongst the off-diagonal elements are preserved. If the diagonal element δ_i of row i of a confusion matrix is scaled by α_i, it is easy to show that the scaling for each off-diagonal element of row i that preserves the relative probability of on-diagonal and off-diagonal elements in the row is

$$\beta_i = \frac{1 - \alpha_i \delta_i}{1 - \delta_i} \tag{1}$$

provided $0 < \alpha_i \delta_i < 1$. The second technique was based on Moore's work. The consonant confusion data he used was available at SNRs of 12, 6, 0, −6, −12 and −18 dB (at a bandwidth of 200–6500 Hz) and he was able to use this to make predictions for any SNR. Moore used a standard method from "multidimensional scaling", a technique in the field of mathematical psychology described by Duda and Hart as "the process of finding a configuration of points whose inter-point distances correspond to dissimilarities" (Duda et al., 2001). The idea is to transform confusion matrices to distance matrices between the sounds, to scale these distances according to the amount of noise present, and then transform back to a confusion matrix. Moore used an expression due to Wilson (1967) to estimate a distance table from a matrix of confusion counts of the type shown in Fig. 2. The expression is

$$D(i,j) = 0.5 \log_{10} \left| \frac{f(i,i) f(j,j)}{f(i,j) f(j,i)} \right| \tag{2}$$

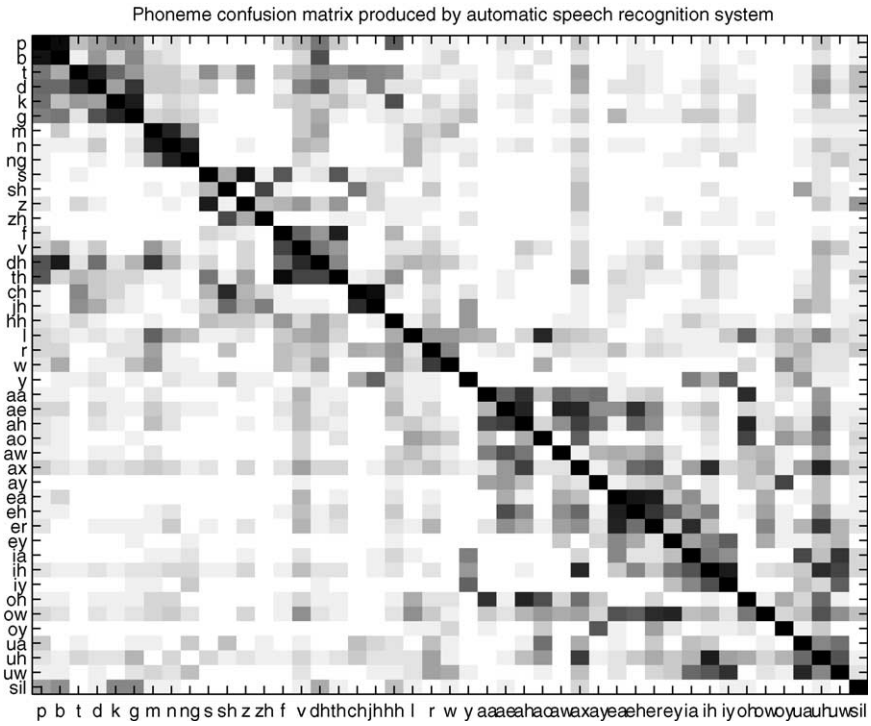

Fig. 2. The phoneme confusion matrix generated by the speech recogniser.

where $f(i,j)$ is an entry in the matrix of counts. The assumption is then made that the addition of noise decreases the distances between sounds uniformly, so that a new matrix

$$D'(i,j) = \eta D(i,j) \quad \forall i,j \tag{3}$$

is constructed, where η is the "noise-figure". D' is then transformed back into a new confusion matrix $C'(i,j)$ using

$$C'(i,j) = \frac{w_j 10^{-2D'(i,j)}}{\sum_{k=1}^{P} w_k 10^{-2D'(i,k)}} \tag{4}$$

In Eq. (4) P is the number of phonemes and w_i is a vector of response weights or biases which is to account for asymmetry in the original f matrix. In general $f(i,j) \neq f(j,i)$, but since $D(\ )$ is a distance function, it must be the case that $D(i,j) = D(j,i)$, and Eq. (4) ensures that this is the case. The asymmetry in the matrix $f(\ )$ is caused mainly by the fact that different phonemes have different attributes which cause them to be more or less likely to be recognised. The expression given by Shephard (1957) for this vector w of weights is

$$w(j) = \frac{P \sum_{i=1}^{P} \sqrt{\frac{f(i,j)}{f(i,i)}} \Big/ \sqrt{\frac{f(j,j)}{f(j,i)}}}{\sum_{k=1}^{P} \sum_{i=1}^{P} \sqrt{\frac{f(i,k)}{f(i,i)}} \Big/ \sqrt{\frac{f(k,k)}{f(k,i)}}} \tag{5}$$

The more symmetrical a confusion matrix is, the closer these weights are to 1.0.

3.6. Modelling spontaneous speech effects

Some effects of assimilation, insertion and deletion of phonemes in spontaneous or rapid speech were modelled by manually modifying the pronunciation strings corresponding to the phrases in the phrase-set. These modifications were done in consultation with a phonetician and were intended to model likely effects in rapid speech. We group these effects under the term "co-articulation", although strictly, co-articulation occurs in all speech, even carefully articulated speech. The following modifications were made:

1. Any `t` at the end of a phrase was removed (e.g. FOUR-EIGHT = `/f ao ey t/` → `/f ao ey/`)
2. A double `n` within a phrase was deleted (e.g. SEVEN-NINE = `/s eh v n n ay n/` → `/s eh v n ay n/`)
3. A double `s` within a phrase was deleted (e.g. SIX-SEVEN = `/s ih k s s eh v n/` → `/s ih k s eh v n/`)
4. A double `f` within a phrase was deleted (e.g. GOLF-FOUR = `/g oh l f f ao/` → `/g oh l f ao/`)
5. A double `t` within a phrase was deleted (e.g. EIGHT-TWO = `/ey t t uw/` → `/ey t uw/`)
6. Any `t` preceding an `h` was deleted (e.g. EIGHT-THREE = `/ey t t th r iy/` → `/ey th r iy/`)
7. Any `k` preceding a `t` was deleted (e.g. QUEBEC-TWO = `/k w ih b eh k t uw/` → `/k w ih b eh t uw/`)
8. Any `t` preceding an `s` was deleted (e.g. EIGHT-SIX = `/ey t s ih k s/` → `/ey s ih k s/`)
9. Any `t` preceding a `z` was deleted (e.g. FOXTROT-ZERO = `/f oh k s t r oh t z ia r ow/` → `/oh k s t r oh z ia r ow/`)
10. Any `t` preceding a `n` was deleted (e.g. EIGHT-NINE = `/ey t n ay n/` → `/ey n ay n/`)
11. Any `/v` preceding an `f` was deleted (e.g. FIVE-FOUR = `/f ay v f ao/` → `/f ay f ao/`)
12. Any `z` following an `s` was deleted (e.g. SIX-ZERO = `/s ih k s z ia r ow/` → `/s ih k s ia r ow/`)
13. Any `f` following a `t` was deleted (e.g. EIGHT-FOUR = `/ey t f ao/` → `/ey f ao/`)
14. Any trailing `r` was deleted (e.g. OSCAR = `/oh s k ax r/` → `/oh s k ax/`).

Some of these modifications are accurate representations of pronunciations in spontaneous speech (for instance, the removal of any re-occurring sound, rules 2–5). Some are over-simplifications, notably those that concern the removal of the final `t` in a word (rules 6, 8–10). This reduces the word EIGHT to the single vowel `y` i.e. the word `/A/`, and `/A-SIX/` is not a very realistic realisation of the phrase EIGHT-SIX even in very rapid speech. However, the object of this part of the study was to model mis-classification for the worst case (i.e. the most rapid and co-articulated

speech) and so these possibly somewhat unrealistic pronunciations were included.

4. Speech data and processing

4.1. The speech recogniser

The speech recogniser used to provide the confusion matrices used in these experiments was a hidden Markov model recogniser available as part of the Entropic Hidden Markov Model Toolkit (HTK) software, version 2.2 (Jansen et al., 1996). The speech recogniser was trained using speech data from the WSJCAM0 database (Fransen et al., 1994). This database was collected at the Cambridge University Engineering Department in 1994 and consists of sentences read from the Wall Street Journal newspaper by 53 males and 39 females with British English accents. The recording quality is high: recording was done in a soundproof room using a Sennheiser HMD414-6 close-talking microphone at a 16 kHz sampling-rate and using 16-bit sample resolution. After recording, each sentence waveform was segmented at both the word and the phoneme level using an automatic procedure. About 90 sentences from each of the speakers in the WSJCAM0 database (a total of approximately 12 h of speech) was used to train a set of 45 phoneme models. The speech waveforms were first filtered using a bandpass filter in the range 300–2500 Hz to simulate the restricted bandwidth of the radio telephone over which the speech is passed. The upper bandwidth limit of 2500 Hz is very low for speech communication. The reason for this is that the region above 2500 Hz is used for signalling information (e.g. press-to-talk signals), radio control and monitoring. The waveforms were then converted to a mel-frequency cepstral coefficient (MFCC) representation (Davis and Mermelstein, 1980), which consisted of 12 MFCCs and a log-energy value, together with the first and second differentials of these (39 components in all). Each phoneme model consisted of a three-state left-to-right HMM, with a five-mode 39-d Gaussian mixture modelling the distribution of vectors in each state. A diagonal covariance matrix was used for each component of the distribution.

4.2. Generation of a confusion matrix

To generate the confusion matrix, the recogniser was configured to output the sequence of phonemes that best matched the input speech, unconstrained by the need to form words or sequences of words. By turning off these constraints, a confusion matrix is obtained which depends only on acoustic confusion performance and not on word or language context. The confusion matrix was generated by recognising the same speech as was used to train the recogniser, and using DP to align the transcription and recognition strings. Insertions (extra phonemes in the recognised string not present in the transcription) were disregarded. It is considered poor practice to test on the training material when quoting recognition accuracy results because this approach overestimates the performance of the recogniser on unseen data. However, in this case, our goal was to generate a confusion matrix rather than to measure recognition performance. The normalised phoneme confusion matrix is shown in Fig. 2. In this matrix, the phonemes have been arranged in groups, with the main division being between consonants (upper) and vowels (lower). In Fig. 2, as in all other confusion matrices in the paper, probability has been non-linearly coded on a grey scale to deliberately emphasise low probabilities. Accordingly, all probabilities above 0.45 are black.

Most of the dark colouring in Fig. 2 is concentrated on the diagonal of the matrix indicating that correct recognitions predominate—the phoneme accuracy of the recogniser is 55.0%. It can also be seen that there appear to be two square areas in which mis-recognitions occur, around the upper diagonal (consonants) and the lower (vowels). The fact that there are few mis-recognitions outside these two squares indicates that vowels and consonants are rarely confused by the recogniser.

Although there are single squares of grey scattered about these two squares, there is evidence of a pattern to the mis-recognitions, especially for the consonants. The consonants (`b` through `zh`) have been grouped according to their manner of articulation: stops, nasals, fricatives, laterals. Consonants within a group are similar in their

articulation and acoustic characteristics and hence more likely to be confused by humans. This behaviour is shown in the machine confusion matrix by the presence of squares of grey colouring around the diagonal associated with these groups. The vowels (aa to uw) have been sorted alphabetically in Fig. 2, but because of the close relationship between orthography and sound, this reflects their "closeness" in phonetic space and a similar pattern of confusion near to the diagonal is seen. The question of how similar the confusion matrix generated by the recogniser is to a human confusion matrix is taken up in detail in Section 5.2.

5. Experimental details and results

5.1. The words and vocabularies used in the experiments

There appears to be no standard vocabulary or syntax for an aircraft call-sign. In a circular to airline operators (Services, 1996), the CAA advises them to "avoid use of similar numerical call-signs within the same company", "avoid multiple use of the same digit", "consider a balance of alphanumeric and numeric call-signs". This study has focused on the digits and the "airline alphabet" (ALPHA, BRAVO, ...,ZULU). Table 4 gives the list of words used in the study together with their pronunciations. Notice that four words have two pronunciations. Three of these words (FOUR, OSCAR, VICTOR) are words that end with r and can be pronounced with or without the final r. In addition, PAPA can be pronounced as /p ax p aa/ or as /p ae p ax/ and SIERRA has two variants depending on how one pronounces the central diphthong. The BEEP (British English Example Pronouncing) dictionary (Robinson et al., 1996) was used to look up the pronunciations for each word used in the study. Each pronunciation was checked and in one case (SIERRA) altered. The single pronunciation ZERO was used for "0", and OH and NOUGHT were not included.

The confusion performance of two phrase-sets was investigated in these experiments:

1. the digit-triple (DT) phrase phrase-set (ONE-ONE-ONE, ONE-ONE-TWO, ...,ZERO-ZERO-ZERO);
2. the alphabet-digit (AD) phrase phrase-set (ALPHA-ONE, ALPHA-TWO, ...,ZULU-ZERO).

The DT phrase-set was made by first generating all possible triples of the eleven digit words ($11\times11\times11=1331$ phrases). The pronunciations of the three words in a triple were then concatenated into a string for use by the DP algorithm. A set of co-articulated digit-triple pronunciations (CDT) was made by editing the pronunciation strings according to the rules described in Section 3.6. The AD and CAD phrase vocabularies were made in the same way. The AD phrase-set had $31\times11=341$ phrases when no co-articulation was modelled and $29\times10=290$ phrases when co-articulation was modelled (some alternative pronunciations disappear with co-articulation modelling). The vocabularies are summarised in Table 5.

Vocabularies were tested at particular SNRs by simulating the effect on the confusion matrix of additive broadband noise. Two techniques were tested for this: direct scaling of the confusion matrix values and transformation of distances (as discussed in Section 3.5).

5.2. Experimental testing of simulation of different SNRs

Section 3.2 outlines the reasons why it was decided to use a machine-generated confusion matrix rather than matrices derived from testing of humans. Although it was stated in Section 4.2 that the machine-generated matrix was similar in the pattern of its confusions to the human matrix, no evidence was offered for this. In addition, it was not known what the Human Equivalent Noise Ratio (HENR, Section 2.1) of the recogniser was. In order to check that the machine did indeed exhibit a pattern of mis-recognitions similar to those made by a human and also to calibrate it in terms of human performance, the confusions for the consonants that were studied in Miller and

Table 4
The words used in the study and their pronunciations

ONE	w ah n	KILO	k iy l ow
TWO	t uw	LIMA	l iy m ax
THREE	th r iy	MIKE	m ay k
FOUR_1	f ao	NOVEMBER	n ow v eh m b ax
FOUR_2	f ao r	OSCAR_1	oh s k ax
FIVE	f ay v	OSCAR_2	oh s k ax r
SIX	s ih k s	PAPA_1	p ax p aa
SEVEN	s eh v n	PAPA_2	p ae p ax
EIGHT	ey t	QUEBEC_1	k w ih b eh k
NINE	n ay n	QUEBEC_2	k w ax b eh k
ZERO	z ia r ow	ROMEO	r ow m iy ow
ALPHA	ae l f ax	SIERRA_1	s ia r ax
BRAVO	b r aa v ow	SIERRA_2	s ih ea r ax
CHARLIE	ch aa l iy	TANGO	t ae ng g ow
DELTA	d eh l t ax	UNIFORM	y uw n ih f ao m
ECHO	eh k ow	VICTOR_1	v ih k t ax
FOXTROT	f oh k s t r oh t	VICTOR_2	v ih k t ax r
GOLF	g oh l f	WHISKEY	w ih s k iy
HOTEL	hh ow t eh l	XRAY	eh k s r ey
INDIA	ih n d ia	YANKEE	y ae ng k iy
JULIET	jh uh l ih eh t	ZULU	z uw l uw

Table 5
The vocabularies used in this study

Phrase-Set name	Example	Coarticulation modelled?	Number of phrases
DT	TWO-EIGHT-FOUR	No	1331
CDT	TWO-EIGHT-FOUR	Yes	1000
AD	FOXTROT-ZERO	No	341
CAD	FOXTROT-ZERO	Yes	290

Nicely's paper were extracted from the machine-generated matrix and normalised to probabilities. The resulting matrix was transformed to simulate the effect of added noise using the two methods for transforming confusion matrices discussed in Section 3.5 (i.e. direct scaling of the values and transformation of distances), and compared with the six human confusion matrices (normalised to probabilities) in Miller and Nicely's paper. These matrices were made at SNRs of 12 dB, 6 dB, 0 dB, −6 dB, −12 dB and −18 dB at a bandwidth of 200–6500 Hz (Tables I–VI in Miller and Nicely's paper). The two matrices were compared by summing the squared differences between equivalent elements, and normalising this sum to generate a stress figure between 0 and 1 (Kruskal, 1964). The expression for the stress σ between two $N \times N$ matrices A and B is given in Eq. (6).

$$\sigma = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (A(i,j) - B(i,j))^2}{\sum_{i=1}^{N} \sum_{j=1}^{N} A^2(i,j) \sum_{i=1}^{N} \sum_{j=1}^{N} B^2(i,j)} \quad (6)$$

For each human confusion matrix, a transformation of the machine-matrix was sought which minimised the stress between the two matrices. This was accomplished by a search though values of α (for the scaling method) and η (for the distance transformation method) to find the transformation that gave minimum stress. To enable interpretation of the stress value, several (30) randomly generated confusion matrices were made for each SNR, and the mean and standard deviation of the stress value between these matrices and the appropriate human confusion matrix was calculated. These matrices were generated by setting the diagonal value for a row to be the same as the

Table 6
Results of experiment comparing human confusion matrices with transformed machine-generated confusion matrices and randomly generated confusion matrices

SNR (dB)	Direct scaling of machine generated matrix		Distance transformation of machine-generated matrix		Randomly-generated matrix
	Optimum value of α	Stress	Optimum value of η	Stress	Average stress
12	1.25	0.002	1.57	0.001	0.0014 ± 0.0002
6	1.16	0.004	1.34	0.003	0.005 ± 0.0007
0	1.08	0.009	1.09	0.01	0.016 ± 0.0017
−6	0.85	0.04	0.76	0.04	0.09 ± 0.007
−12	0.62	0.14	0.53	0.12	0.29 ± 0.012
−18	0.31	0.59	0.10	0.19	0.77 ± 0.019

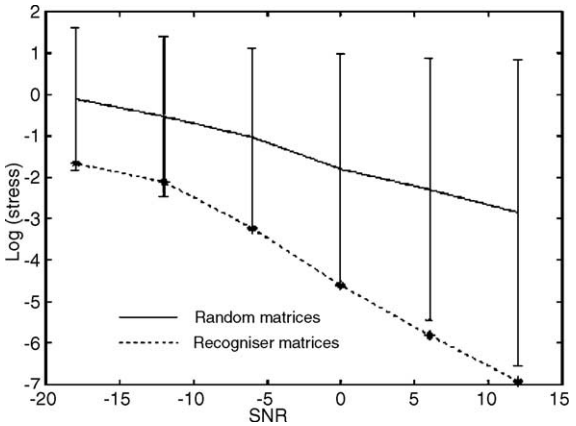

Fig. 3. The (log) stress between confusion matrices generated by the speech recogniser and from 30 "randomly" generated matrices. Error-bars are $\pm 2\sigma$.

diagonal value of the human matrix, and then redistributing the remaining probability for the row randomly amongst the off-diagonal elements. This has the effect of generating a confusion matrix that has the same overall accuracy as a human confusion matrix (since the diagonal elements are the same) but a random pattern of errors. The results of these experiments on stress are given in Table 6.

The following observations can be drawn from Table 6 and Fig. 3.

1. The stress between both the machine-generated and the randomly generated matrices and the human matrices increases as the SNR decreases. This is as expected, since at higher SNRs, the accuracy is high, so most of the probability is concentrated on the diagonal, and hence most off-diagonal entries will be close to zero.
2. For 0 dB SNR and below, both techniques for transforming machine-generated matrices give matrices that are statistically significantly lower in stress than randomly generated matrices, indicating that they are more similar to human confusion matrices. In fact the distance-transformation technique achieves a statistically significant difference from random at an SNR of 6 dB.
3. There is little to choose between the two techniques for transforming machine-generated matrices at high SNRs, but the distance-transformation technique gives a much lower stress value at −18 dB SNR.
4. It is interesting that both techniques have a scaling close to 1.0 for a 0 dB SNR. For both techniques, a scaling of 1.0 leaves the matrix unaffected, indicating that the HENR of the recogniser used is in the region of 0 dB (i.e. the speech recogniser has performance approximately equivalent to a human listening at 0 dB SNR).

This experiment indicated that transforming machine-generated confusion matrices could provide matrices that approximated human performance. It should be borne in mind that the bandwidth of the speech used to make the machine-generated confusion matrix (300–2500 Hz) was significantly lower than the bandwidth used by Miller and Nicely (200–6500 Hz). The more restricted bandwidth of the speech used by the recogniser would have the effect of lowering the

recogniser's accuracy and altering the pattern of mis-recognitions it made. It is not possible to say whether the recogniser confusion matrix made at 300–2500 Hz bandwidth is actually closer to a human confusion matrix at this bandwidth than the human confusion matrices in Miller and Nicely's paper that were made at a higher bandwidth. In view of the slightly superior stress performance of the distance-transformation technique over the scaling technique, it was decided to use the former for simulating different SNRs in subsequent experiments. In Figs. 4 and 5, the human confusion matrices from Miller and Nicely's paper are shown together with the optimally transformed machine-generated matrix using the distance transformation method.

Comparing the human and machine confusion matrices, it can be seen that the patterns of the blocks of confusion around the diagonals are similar at all SNRs. However, as the SNR decreases, in the machine confusion matrices, bands become more and more prominent on either side of the diagonal whereas the human matrices become more random in their pattern of confusions. These bands are due to confusion of voiced and unvoiced consonants in the machine. This effect does occur in the human confusion matrices (faint bands are discernible) but it is not nearly so marked.

5.3. Simulation results

In this section, we summarise the results of experiments aimed at comparing the confusability of the two vocabularies at different SNRs. The six SNRs used by Miller and Nicely (12, 6, 0, −6, −12 and −18 dB) were suitable points at which to conduct our experiments, as we had experimented with transforming our machine confusion matrix to work at these points. Simulations of the confusions for the four vocabularies (DT, CDT, AD and CAD) were run at all six SNRs. Examination of the results for SNRs of 6 and −6 dB showed that results at these SNRs were very similar to the results for 12 and 0 dB respectively, so they have been omitted from the presentation of results.

The resulting phrase-set confusion matrices from the different vocabularies and SNRs were analysed in three different ways:

1. comparison of the predicted accuracies (Section 5.3.1);
2. comparison of the distributions of a statistic derived from the confusion matrices (Section 5.3.2);
3. estimating a "potential confusion matrix" for the individual words within a phrase by identification of the "closest" phrase to the stimulus phrase (Sections 5.3.3 and 5.3.4).

5.3.1. Phrase-set accuracy

The accuracy predicted by the model for a given vocabulary at a given SNR is estimated by computing the probability of each response-phrase given each stimulus-phrase, and then noting the number of "correct recognitions". A "correct recognition" occurs when, for a given stimulus phrase, the response phrase whose probability is highest is the stimulus phrase. The raw data from a simulation of a phrase-set consist of six confusion matrices (one for each SNR) each of approximate size 1000×1000. We have attempted to draw summary data and salient points about the confusions from this very large body of data.

The mean accuracies predicted by the model are given in Table 7. The trend of the accuracies shown here is typical of the intelligibility to humans of speech in noise, in that intelligibility remains high until a "cut-off" is reached (here between −12 dB and −18 dB) and then falls off sharply. On a given row in Table 7, the accuracy figure in a given column is either the same or lower than the accuracy in the column to its left, which means that the predicted recognition accuracy never increases as the SNR is decreased. This, of course, is a minimal requirement for a model to be considered realistic. We are sceptical about the *absolute* values of the figures in Table 7 because, given the many assumptions that have been made in the simulations, the ability of the model to predict accurately recognition accuracy at a given SNR must be doubted. However, it should be possible to compare and rank *relative* confusability effects, both within a certain phrase-set and across different vocabularies, and we concentrate on this kind of analysis here.

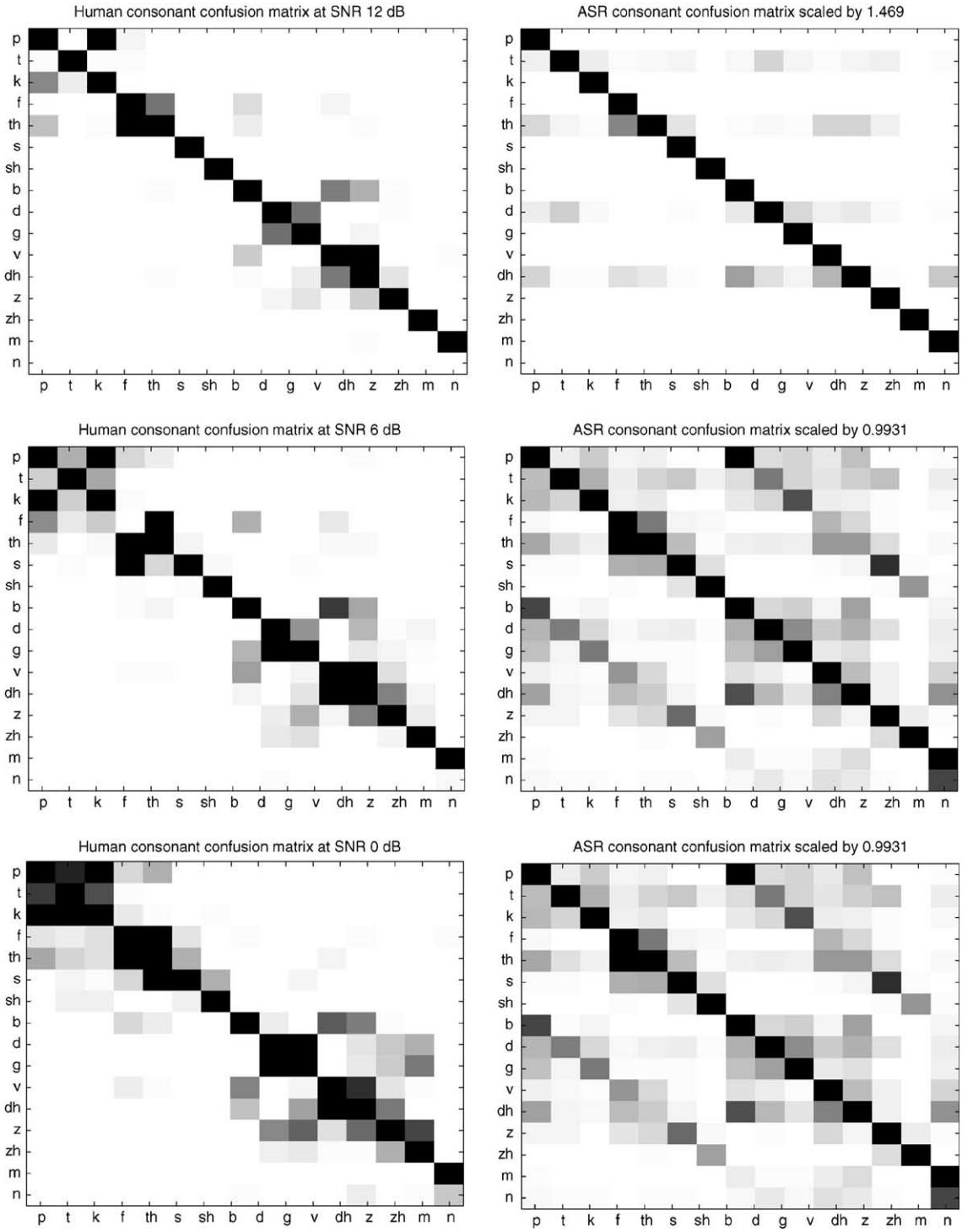

Fig. 4. Comparison of human confusion matrices (left column) and transformed machine-generated matrices (right column) at 12 dB (top), 6 dB (middle) and 0 dB (bottom).

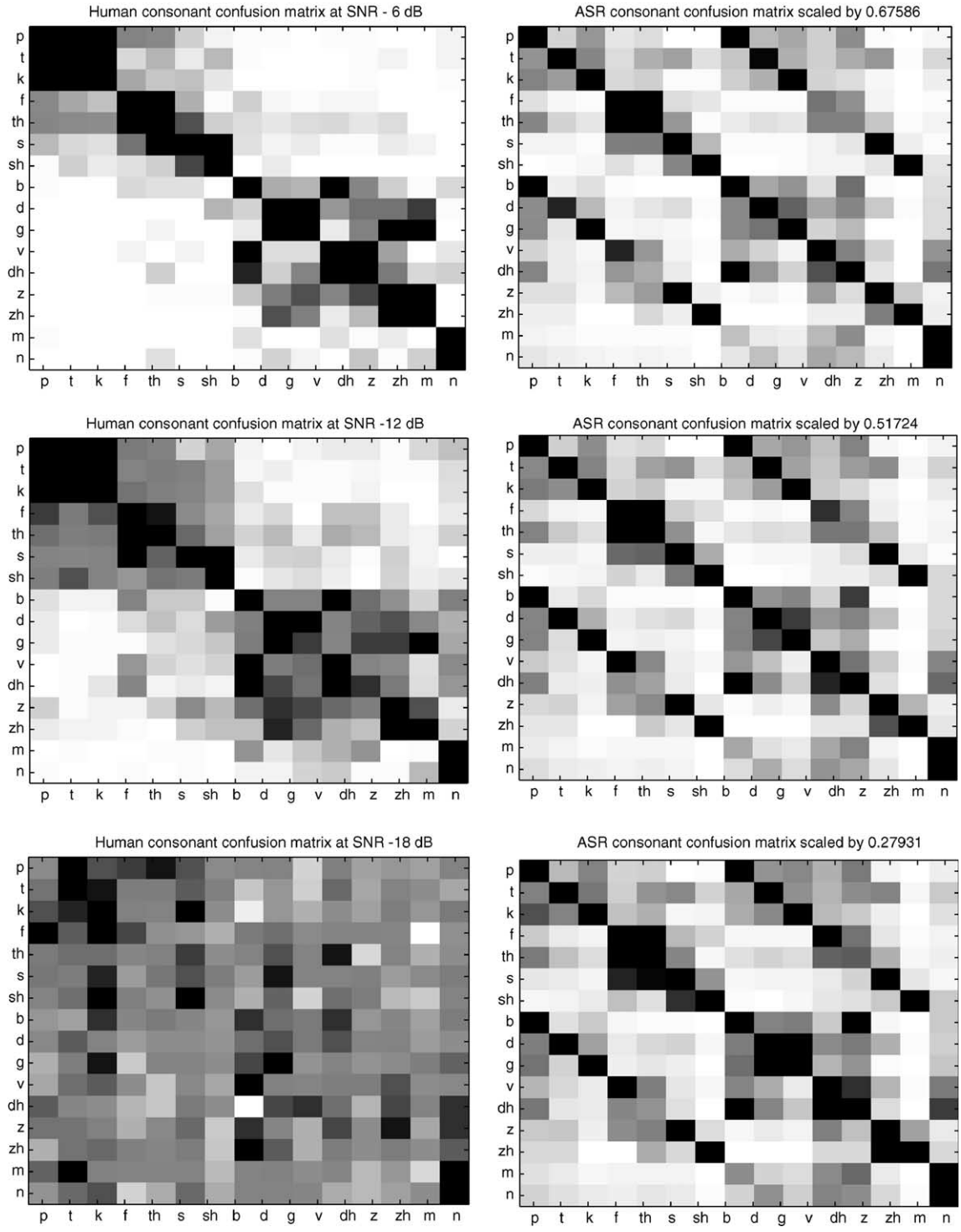

Fig. 5. Comparison of human confusion matrices (left column) and transformed machine-generated matrices (right column) at −6 dB (top), −12 dB (middle) and −18 dB (bottom).

Table 7
Accuracies predicted by the model for the four vocabularies

Phrase-set	No of phrases	12 dB	0 dB	−12 dB	−18 dB
DT	1331	99.99	99.98	96.6	4.44
CDT	1000	99.98	99.91	95.72	4.06
AD	341	99.99	99.99	99.99	9.85
CAD	290	99.99	99.99	99.76	10.06

5.3.2. The random variable δ as a confusion indicator

In cases where the probabilities from the correct response phrase are close to those from one or more incorrect response phrases, a small alteration to the probabilities can have a large effect on the overall average accuracy. The model makes several assumptions which make the probabilities in any confusion matrix it generates subject to errors. Hence the estimated mean accuracy may not be a very reliable guide to actual performance of a vocabulary at a certain SNR. A more qualitative but potentially more insightful measurement of the relative potential confusability of two vocabularies V_1 and V_2 is to compare the *distributions* of the random variables δ_{V_1} and δ_{V_2}, where

$$\begin{aligned}\delta_{V_i} \in \{&C_i(1,1) - C_i(1,2), C_i(1,1) - C_i(1,3), \ldots,\\ &C_i(1,1) - C_i(1,N_i), C_i(2,2) - C_i(2,1),\\ &C_i(2,2) - C_i(2,3), \ldots, C_i(2,2) - C_i(2,N_i),\\ &\vdots \qquad\qquad \vdots\\ &C_i(N_i,N_i) - C_i(N_i,1), C_i(N_i,N_i) - C_i(N_i,2), \ldots,\\ &C_i(N_i,N_i) - C_i(N_i,N_{i-1})\} \end{aligned} \tag{7}$$

In Eq. (7), $C_i(j,k)$ is the phrase-set confusion matrix for phrase-set V_i and N_i is the number of phrase-set items, so that δ_{V_i} is the set of differences between the on-diagonal probability (from the correct response) and the off-diagonal probabilities (from the incorrect responses) for each row. Clearly, the closer this value is to zero, the more likelihood there is of confusion. It is interesting to compare the distributions of δ_V for the two vocabularies at the same SNR. A priori, there are two effects that should make the AD phrase-set easier to recognise than the DT phrase-set:

1. The AD phrase-set is about 1/4 of the size of the DT phrase-set. Hence if the vocabularies were otherwise similar, AD would have a lower confusion probability than DT.
2. The AD phrase-set uses many more different phonemes than the DT phrase-set because of the presence of the "phonetically rich" airline alphabet words.

There is also one effect that should make it harder to recognise the AD phrase-set: the AD phrase-set has an average of 7.9 phonemes in a phrase whereas the DT phrase-set has an average of 9.0, so AD would be slightly harder to recognise if the phrase-set content were similar.

Fig. 6 shows the values of δ_V for the AD phrase-set (top) and the DT phrase-set (bottom) for an SNR of −18 dB. It can be seen that the distribution for the AD phrase-set has a median that is greater than that for the DT phrase-set and has a long tail of high values. If we can make the reasonable assumption that two phrase-set items v_i and v_j are likely to be confused by a listener if the value of $C(i,j)$ is less than some threshold value T (T must be determined experimentally), then the interpretation of Fig. 6 is that the AD phrase-set is inherently less confusable than the DT. This is backed up by the results in Table 7 for −18 dB SNR, where accuracy on the AD phrase-set is double that on the DT phrase-set.

Fig. 7 shows the values of δ_V for the CDT phrase-set (top) and the DT phrase-set (bottom).

The distributions showing that the effect of co-articulation on the probability distributions appears to be very small and we would not expect much difference in confusability between these two vocabularies. The same effect is observed in comparing the AD and CAD vocabularies i.e. there is little difference in relative confusion probabilities caused by the co-articulation modelling.

5.3.3. Analysis of confusions of individual words in the digit triples (DT) phrase-set

Potential confusions for *individual* words within the DT phrase-set at a given SNR were identified as follows:

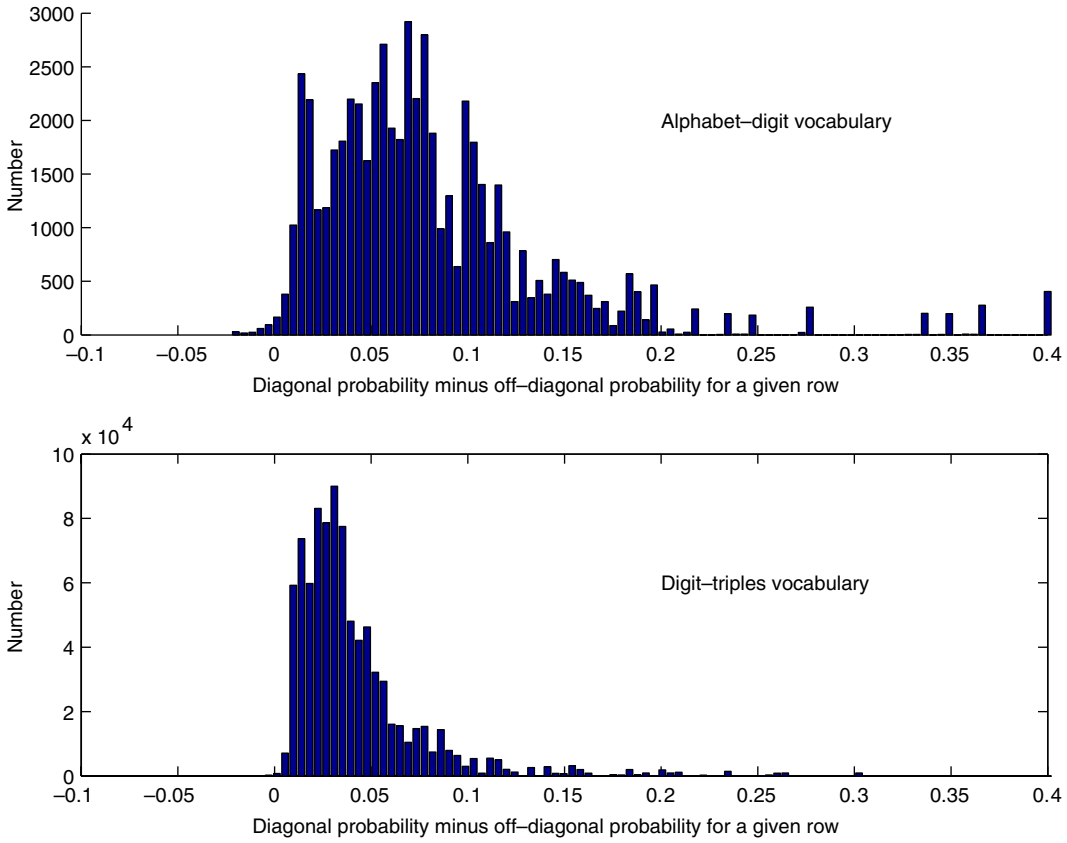

Fig. 6. Comparison of differences in (on-diagonal probability)−(off-diagonal probability) for AD phrase-set (top) and DT phrase-set (bottom).

1. for each stimulus phrase, zero the (diagonal) probability associated with a correct recognition of this phrase and hence identify the most probable *incorrect* phrase;
2. compare the stimulus phrase with the the most probable incorrect phrase to form a "potential confusion matrix", *PCM*, for a given SNR (for instance, if the stimulus string were ONE-TWO-THREE and the most probable incorrect response ONE-TWO-FIVE, the element $PCM(3,5)$ would be incremented);
3. when all the phrases have been processed, normalise *PCM* across its rows to form probabilities.

This analysis is based on identifying the single "closest" phrase to the stimulus phrase rather than examining the complete distribution of probabilities of responses as was done in Section 5.3.1. The rationale for this approach is that we believe that the absolute probabilities in a predicted confusion matrix C are subject to error, but expect the *ranking* of responses associated with a given stimulus phrase to be more robust. Examining the most probable incorrect phrase gives insight into potential mis-recognitions in the phrase-set: a "potential confusion matrix" shows the most probable confusions for a stimulus phrase when the SNR drops sufficiently low for confusions to

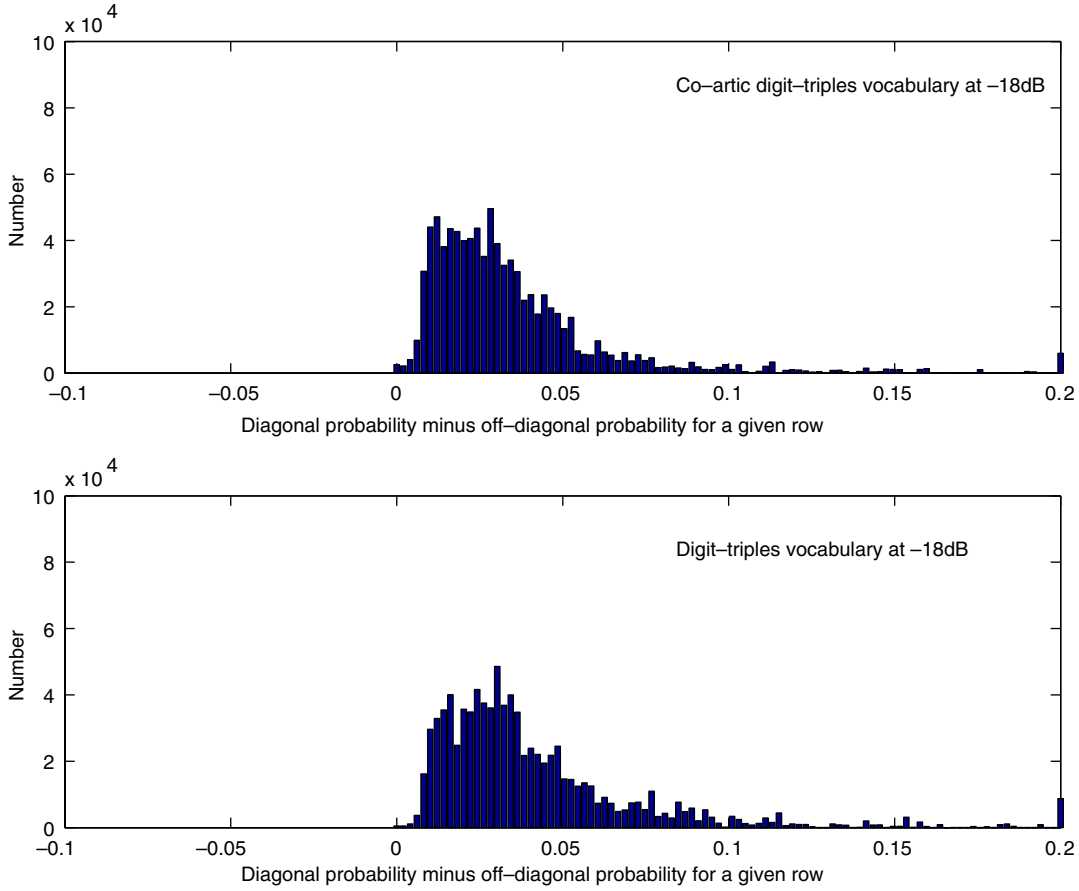

Fig. 7. Comparison of differences in (on-diagonal probability)–(off-diagonal probability) for CDT phrase-set (top) and DT phrase-set (bottom).

occur for listeners. Table 8 shows the PCM matrix for the DT phrase-set at 12 dB SNR.

The model predicts some well-known confusions in the digit phrase-set such as ONE/NINE, FOUR/FIVE and TWO/EIGHT, although not FIVE/NINE. The least confusable digits are predicted to be ZERO, SIX and THREE and the most confusable ONE and NINE.

This approach was refined to examine whether the mis-recognition rate was influenced by the position of a digit in the phrase. Three separate PCM matrices were estimated, one for each digit position. The matrices are not given here for reasons of space, but the average "accuracy" in each digit position was as follows: first digit = 83.9%, central digit = 52.3%, final digit = 81.4%. If the null hypothesis is that each digit position should have the average "accuracy" (72.5%) within sampling error, this hypothesis can be rejected at the 0.1% level for the central digit. There is significant difference between accuracy for the outer digits. Hence the model predicts that the central digit is significantly more likely to be mis-recognised than the outer digits.

Table 8
Potential confusion matrix for DT phrase-set at 12 dB SNR

	ONE	TWO	THREE	FOUR_1	FOUR_2	FIVE	SIX	SEVEN	EIGHT	NINE	ZERO
ONE	0.143	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.857	0.000
TWO	0.000	0.672	0.000	0.129	0.000	0.000	0.000	0.000	0.198	0.000	0.000
THREE	0.000	0.094	0.904	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000
FOUR_1	0.000	0.088	0.000	0.625	0.000	0.287	0.000	0.000	0.000	0.000	0.000
FOUR_2	0.000	0.000	0.000	0.003	0.576	0.421	0.000	0.000	0.000	0.000	0.000
FIVE	0.000	0.000	0.000	0.303	0.198	0.499	0.000	0.000	0.000	0.000	0.000
SIX	0.000	0.006	0.000	0.000	0.000	0.000	0.981	0.003	0.011	0.000	0.000
SEVEN	0.074	0.000	0.000	0.000	0.000	0.000	0.000	0.923	0.003	0.000	0.000
EIGHT	0.000	0.273	0.003	0.003	0.000	0.000	0.008	0.003	0.711	0.000	0.000
NINE	0.686	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.314	0.000
ZERO	0.000	0.000	0.017	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.983

Table 9
Analysis of potential confusions in the AD phrase-set

SNR (dB)	# alphabet confusions	# digit confusions	Mean alphabet probability	Mean digit probability
12	30	311	1.00×10^{-6}	6.8×10^{-5}
0	43	298	3.92×10^{-6}	11.9×10^{-5}
−12	43	298	0.0037	0.0073
−18	140	201	0.053	0.049

Table 10
Most common potential confusions in the AD phrase-set

Alphabet words	Digit words
ALPHA/DELTA	ONE/NINE
DELTA/VICTOR	TWO/FOUR
ECHO/X-RAY	TWO/EIGHT
KILO/LIMA	THREE/TWO
OSCAR/ECHO	FIVE/NINE
TANGO/YANKEE	
ZULU/KILO	

5.3.4. Analysis of confusions of individual words in the alphadigit (AD) phrase-set

Potential confusions within the AD phrase-set at a given SNR were identified using the same technique described in Section 5.3.3. In this case, the highest probability response phrase was compared with the stimulus phrase to find whether the "mis-recognition" was in the alphabet word or the digit word (there were no cases where the highest probability response differed from the stimulus in both words). The mean probability of the highest response phrase was also computed for the case when the mis-recognition was an alphabet word, and the case when it was a digit.

Table 9 shows the results of this analysis.

It can be seen that the confused word was about 8–10 times more likely to be a digit than an alphabet word for SNRs of 12, 0 and −12 dB, and the associated probability of confusion was 2–50 times higher in cases where the confused word was a digit rather than an alphabet word. At −18 dB, the numbers of potential confusions and associated probabilities become more similar for the two groups of words. The most common confusions of the two groups are given in Table 10.

When all 340×340 incorrect response probabilities were sorted, it was found that in the highest 1000 incorrect probabilities, there were only 44 that were due to "mis-recognition" of an alphabet word rather than a digit. These findings all point to the same conclusion, that the digits are inherently more confusable than the airline alphabet.

5.4. Benchmarking the model performance using real confusions

The performance of the model was benchmarked using a set of confusions available from the CAA database (N.B. none of these confusions led to an air-safety incident). This database con-

tained 61 confusions of digit triples with other triples in which the response was not a transposed version of the stimulus, indicating a perceptual rather than a cognitive confusion. Of these, 59 had a single digit in a certain position different in stimulus and response, and these were used in this evaluation. The performance of the model was evaluated using the DT phrase-set at 12 dB SNR (no information about the SNR under which these confusions were observed was available—it is most likely they were made at a range of SNRs). Each of the 59 "stimulus" phrases (i.e. the digit triple phrases that were spoken), were input into the model and the response phrases produced were ordered by probability. In addition, the rank of the phrase that had been erroneously "recognised" was recorded. A perfect model of mis-recognition would record a rank of one for every example, and a model that simply guessed the answer would give a uniform distribution of ranks with a mean rank close to $999/2 \approx 500$, since there are 999 possible incorrect responses to the stimulus phrase (the correct response was not included in the ordering).

Using these data, the model was tested under two conditions:

1. the possible responses to the stimulus phrase were unrestricted and could be any of the 999 digit triples that were different from the stimulus;
2. the possible responses were restricted to the 27 responses that differ from the stimulus phrase by one digit in one position.

The rationale for testing a model in which the response was restricted to digit triples that differed by only a single digit in one position is that this response is the most likely mis-recognition, as borne out by the examples in the database (59 of the 61 examples showed this pattern of response—the other two examples had two errors). A useful model would be able to give a better prediction of confusion than choosing randomly a response that differs by only a single digit in one position from the stimulus phrase. The results are given in Table 11.

Table 11
Results of comparison of model predictions with real confusions

Model	Mean rank of actual response in model	Lowest rank of actual response
999 possible responses (unrestricted)	17.9	53
27 possible responses	13.2	27

For the unrestricted response case, a random selection for the rank of the actual response would yield a mean rank of about $999/2 \approx 500$. The fact that the mean ranking from the model is 17.9 shows that it is much better than random. This inference is supported by the fact that lowest rank assigned to any one actual response is 53. However, when only 27 responses are allowed, the mean rank obtained using random selection would be $27/2 = 13.5$ and the mean rank of 13.2 obtained using the model is not significantly different from this. In one case, the model assigned the lowest possible ranking, 27, to the actual response.

In addition, it was noted that that the number of confusions in each digit position in the supplied data was as follows:

First digit	11
Central digit	35
Final digit	13

A null hypothesis is that the expected number of errors in each digit position is the same and is equal to $59/3 = 19.66$. The results above lead to rejection of the null hypothesis at the 0.1% level i.e. digits in the central position are significantly more likely to be mis-recognised than digits in the outer positions. This result is in agreement with the model prediction stated in Section 5.3.3, that the probability of a confusion in the central digit is higher than a confusion in either the first or final digits. The confusion probabilities predicted from the three PCM matrices estimated in Section 5.3.3 are 0.161, 0.477 and 0.186 for the first, central and final digit respectively. It is interesting that the predicted confusion probability for the central

digit is about three times that of the first and final digits because the error-rate for the central digit in the real confusion data is also about three times the error-rate for the first and final digits. However, the predictive power of the model needs to be verified before it is possible to state whether this is merely a coincidence.

The analysis suggests that, if it is assumed that the confusion of a digit triple is most likely to be another triple that differs in one location, the model performs no better than making a random choice for the response. However, very little information on the confusions in this database was available: the conditions under which the confusion was made, the SNR of the channel, whether the pilot or controller was a native English speaker etc., are all unknown. It would be wise to conclude that the conditions under which the data for this evaluation was gathered were not sufficiently well-defined or well-controlled to lead to any hard conclusions about the performance of the model, except to say that its performance is not unreasonable, in that its average ranking of the actual response was about 18 compared with an average of 500 that would be obtained by random choice. However, a proper evaluation with controlled data under controlled conditions is required before any firm conclusions about the predictive power of the model can be drawn.

6. Summary and discussion

An examination of the potential confusions between short phrases of the kind that are used in the dialogue between a pilot and an air-traffic controller was made. This study has concentrated on modelling perceptual rather than cognitive confusions. The suitability of using established subjective and objective techniques to estimate confusability was reviewed, and these techniques were considered to be unsuitable for different reasons. A technique using a model for prediction of confusability (based on work originally done by Moore and separately by Simons) was proposed, developed and implemented. The technique differed from the work of Moore in that a confusion matrix derived from a speech recogniser was used rather than one derived on studies on human performance, and the validity of this substitution was tested carefully. The technique was used in a study that simulated the effect of the restricted bandwidth of the communication channel, the effect of additive broadband noise on the speech signal and the effect of a spontaneous speaking style. Two different vocabularies were used: triples of digits (DT phrase-set) and an airline alphabet word (ALPHA, BRAVO, CHARLIE etc.) followed by a digit (AD phrase-set). The overall behaviour of the model, as measured using real confusions extracted from the CAA database, is reasonable, and its predictions of confusable words within phrases accord with experience. However, before any more development is done to the model, its predictions need to be tested and validated by subjective tests using a panel of listeners. These tests will not be unwieldy because they will be restricted to testing of predictions made by the model. However, they will be sufficiently general to enable the model to be properly calibrated for different SNRs. A comparison of the results from the model with the actual results will enable us to identify the assumptions in the model that need to be adjusted or corrected to make it more realistic.

This study has only begun to scratch the surface of a complex problem in which perceptual and cognitive effects are intertwined, and further research is required to gain a deeper insight into the reasons for confusion of call-signs. The goal of the research is to be able to understand the confusions well enough to enable design of a tool that would aid air-traffic controllers in assigning call-signs to aircraft to increase reliability of communication and hence air safety. If successful, this tool would be useful in many situations where it is required to design an optimally intelligible set of phrases from a closed phrase-set.

Acknowledgements

This work was sponsored by the UK National Air Traffic Services.

Appendix A. Correspondence between AzRPABET transcription symbols and IPA symbols

	IPA symbol	ARPABET symbol		IPA symbol	ARPABET symbol
Vowels	i	iy	Consonants	p	p
	ɪ	ih		b	b
	e	eh		t	t
	æ	ae		d	d
	u	uw		k	k
	ʊ	uh		g	g
	ʌ	ah		f	f
	ɒ	oh		v	v
	ə	ax		θ	th
	ɜ	er		ð	dh
	ɔ	ao		s	s
	ɑ	aa		z	z
Dipthongs	eɪ	ey		ʃ	sh
	aɪ	ay		ʒ	zh
	ɔɪ	oy		ʧ	ch
	ɑʊ	aw		ʤ	jh
	əʊ	ow		m	m
	ɪə	ia		n	n
	ɛə	ea		ŋ	ng
	ɔə	ua		l	l
	ʊə	ua		r	r
				j	j
				w	w
				h	h

References

Baddeley, A., 1990. Human Memory: Theory and Practice. Erlbaum Associates, Hove.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing 28, 357–366.

Duda, R., Hart, P., Stork, D., 2001. Pattern Classification. John Wiley and Sons, New York.

Fletcher, H., 1953. Speech and Hearing in Communication. Krieger, New York.

Fransen, J., et al., 1994. WSJCAM0 corpus and recording description (Tech. Rep. No. CUED/F-INFENG/TR.192). Cambridge University Engineering Department.

French, N., Steinberg, J., 1947. Factors governing the intelligibility of speech sounds. Journal of the Acoustical Society of America 19 (1), 90–119.

ICAO Manual of Radiotelephony, 1990. Second Edition.

Jansen, J., Odell, J., Ollason, D., Woodland, P., 1996. The HTK book. Entropic Research Laboratories Inc.

Kruskal, J., 1964. Multidimensional scaling by optimising goodness of fit to a nonmetric basis. Psychometrika, 1–27.

Mendel, L. et al., 1998. Speech intelligibility assessment in a helium environment. ii. the speech intelligibility index. Journal of the Acoustical Society of America 104 (3), 1609–1615.

Miller, G., Nicely, P., 1955. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America 27, 338–352.

Moore, R., 1977. Evaluating speech recognisers. IEEE Transactions of the Acoustics, Speech and Signal Processing 25 (2), 176–183.

Peterson, G., Barney, H., 1952. Control methods used in a study of vowels. Journal of the Acoustical Society of America 24 (2), 175–184.

Pickett, J., 1957. Perception of vowels heard in noises of various spectra. Journal of the Acoustical Society of America 29 (5), 613–620.

Robinson, T. et al., 1996. The British English Example Pronunciation (BEEP) dictionary is available from svr-ftp.eng.cam.ac.uk/comp.speech/dictionaries/beep.tar.gz.

Services, N.A.T., 1996. December. CAA United Kingdom aeronautical information circular. Number AIC 112/1996.

Shephard, R., 1957. Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. Pyschometrica 22 (4), 325–345.

Simons, A., 1995. Predictive assessment for isolated-word speaker-independent speech recognisers. In: Proceedings of the Eurospeech. pp. 1465–1467.

Singh, S., Woods, D., Becker, G., 1972. Perceptual structure of 12 American English vowels. Journal of the Acoustical Society of America 52 (6), 1698–1713.

Steeneken, H., Houtgast, T., 1980. A physical method for measuring speech transmission quality. Journal of the Acoustical Society of America 67 (1), 318–326.

Steeneken, H., Houtgast, T., 1985. Description of the rapid speech transfer index (RASTI) method and its foundation. Journal of the Audio Engineering Society 33 (12), 1007.

Steeneken, H., Houtgast, T., 2002. Validation of the revised STI method. Speech Communication 38 (3–4), 412–425.

Vandeelen, G., Blom, J., 1990. Hearing-loss and radiotelephony intelligibility in civilian airline pilots. Aviation Space and Environmental Medicine 61 (1), 52–55.

Wang, M., Bilger, R., 1973. Consonant confusions in noise: a study of perceptual features. Journal of the Acoustical Society of America 54 (5), 1248–1266.

Wilson, K., 1967. Multidimensional analysis of confusion of English consonants. American Journal of Psychology 76, 89–95.