AUTOMATIC ASSESSMENT OF MACHINE TRANSCRIPTIONS

Peter Roach Andrea Dew Paul Rowlands

Department of Linguistics & Phonetics
University of Leeds, U.K.

ABSTRACT

Research in the automatic transcription of speech sounds by computer requires a detailed and accurate comparison between the expert phonetician's transcription and the machine's attempt. A computational technique for assessing the accuracy of a machine transcription is described: differences between segments are expressed in terms of a small number of primitive phonetic features.

INTRODUCTION

A number of modern approaches to the automatic recognition of continuous speech make use of the technique of dividing the stream of speech into a string of segments and labelling these with a chosen set of phonetic category labels ([1], [2], [3], [4], [5]). These categories, which are not necessarily restricted to phoneme-sized units, may be more or less precisely specified. Dalby et al [6] refer to three different types of analysis: Broad Class (identifying segments as, for example, Nasal, Fricative, Vowel), Mid Class (including details such as whether a segment is voiced or not, whether a vowel is front or back, or whether a fricative is strong or weak), and Fine Class, which is roughly equivalent in precision to a phonemic transcription. Given that such techniques have a useful role to play in a speech recognition system, it can be claimed that phonetic science should be able to contribute significantly to their development, both in their design and in the assessment of their performance. This paper deals with the latter application, discussing the extent to which automatic phonetic transcriptions can be accurately evaluated. This is discussed with reference to a system (which we call LUPINS) developed at Leeds University [7] which carries out speaker-independent Broad Class analysis of continuous speech by automatic segmentation and labelling; while the system was developed using a corpus of recordings from 18 speakers, the tests reported below were carried out with new data from new speakers and the system's recognition rules were left unaltered. It is claimed that a computational technique for measuring accuracy as outlined here will make testing much more efficient than a "manual" equivalent [8], and should be valuable in making explicit some of the phonetic principles underlying the analysis.

RECORDING OF ERRORS IN SEGMENTATION AND LABELLING

As explained in Roach et al (op cit), errors in transcription will be of a number of different types: (i) a segment is omitted; (ii) a spurious segment is inserted; (iii) a segment is assigned to the wrong phonetic category; (iv) a segment boundary is located incorrectly on the time axis. All of these errors must be detected and recorded in the assessment procedure, and some score reflecting the level of seriousness of the error must be derived. In our present research work (funded by S.E.R.C./Alvey Grant MMI-053) the assessment is carried out by a computer program which takes a transcription of a passage made by a human expert and compares it with the computer's transcription of the same data. The human transcription is always treated as the correct model (though it sometimes happens that the computer's version causes humans to revise their transcriptions). Since the transcription is typed in in the symbols of the Edinburgh Machine-readable Phonemic Alphabet or the "Alvey" ASCII symbol codes [8], while the computer transcribes using only a very small set of symbols (basically comprising Fricative, Nasal, Vowel, Dip, Stop, Flap, Burst, Silence), it is necessary for the human transcription to be converted into this alphabet before the comparison begins. All segments in both transcriptions are given duration values in csec.

A simple form of assessment was used in our earlier work: each error of types (i) to (iii) above was counted as one error, and a final success rate was arrived at by expressing the total number of errors as a percentage of the total number of segments in the passage. Errors of type (iv) were ignored. Scoring on this basis gave success rates in the region of 80% for informal conversational speech in six different languages with a number of different speakers including female and male. However, it was found that there were many cases where we felt we should treat some errors as "minor" or "forgiveable" (e.g. inserting a very brief Dip (approximant) segment between neighbouring voiced segments, or categorising a sound as a flap when the human had heard it as a brief stop), while other errors were considerably more serious; it was also found that the process of "marking" a machine transcription was a very time-consuming process that needed to be done after each run of LUPINS. It was because of these factors that it was decided to develop an automatic assessment technique. An additional advantage of doing this was that the technique should also make it possible to align an unknown recording of speech with its transcription: this has a number of potential applications in the field of large speech databases.

AUTOMATIC ASSESSMENT OF ACCURACY: EXAMPLES

Two short recorded test passages that were analysed recently are used as examples of the technique. The first passage is by two speakers, one male and one female, and the text is as follows:

M. Hello, operator - operator?

F. Yes, what can I do for you?

M. I'd like to make a telephone call.

The second passage is a male speaker saying "Can you recognise this sentence?". The assessment is done as follows:

(a) The human transcription (H) and the machine transcription (M) are compared symbol by symbol, and each case of matching symbols is scored as one correct symbol.

(b) When a symbol of H is found not to be matched by the corresponding symbol of M, the M transcription is corrected in one of the following ways:

(i) if M has missed a symbol, the symbol from H is inserted, and one error is recorded.

(ii) if M has inserted a symbol that is not pesent in H, that symbol is deleted and one error is recorded.

(iii) if the corresponding M symbol does not match, but subsequent pairs of H and M symbols do match, the M symbol is marked as incorrect, and is replaced by the H symbol. A score for the error between 0 (insignificant) and 1 (complete failure of recognition) is calculated by the procedure described in Section 4 below and added to the errors total.

(c) If the time values of the H segments are known (they are always included in transcription files made within our project, but may be missing from other transcriptions), the time values in M are adjusted to fit them, and the extent of the required adjustment is noted and added to a time-adjustment total score; adjustments in either direction on the time axis are treated as positive numbers. This score is kept separate from the scoring of correct/incorrect segments. Time measurement is done in csec, and the final time-adjustment error score is the number of csec recorded in the time-adjustment total as a percentage of the overall number of csec in the entire passage.

A particular case of a "missed symbol" is found fairly frequently when an intervocalic segment is missed and a very long vowel recorded instead. In the example given below, for example, the H sequence / əɹeɪ / should have been transcribed as VDV , but came out as a long V ; this would result in two errors being recorded, but we feel it is more appropriate to count this as a case of one missed segment.

MEASUREMENT OF ERROR GRAVITY

Our treatment of cases of incorrect symbols in the M transcription is still at a provisional stage, but it is clear that what is needed is some form of distance measure so that a wrong symbol that denotes a segment radically different from the correct one will be counted as nearer the error value 1, and a symbol that is not so different will receive a score that is nearer to zero. We measure distance by comparing segments on a feature by feature basis: in earlier work [8] we used phonetic features based on those of Ladefoged [10], but found difficulties in relating some of the features to our labels (which are essentially defined in acoustic terms) [11]. We are currently working with a set based on those used in the study of perceptual confusions among English consonants by Miller and Nicely [12]: the provisional set of five "primitive" features comprises +/- Voiced; +/- High energy (the term "high" is deliberately ambiguous between "high in amplitude" and "high in frequency", and is used to distinguish / s and ʃ / from other fricatives); +/- Nasal; +/- Transient (non-transient sounds are capable of having an audible steady state, while transients include plosives, bursts, semivowels and flaps) and +/- Fricative. The features could in some cases be given numerical (non-binary) values if wished, but for the purposes of this paper only binary values are used. (It is noticeable that even this small set contains more redundancy than phonologists would approve of). For each feature that was wrong in the M transcription, .2 was added to the overall error score, and the same was added to the "segments correct" total for each feature correctly identified: hence a case of all five features being wrong (e.g. Burst instead of Nasal) would cause 1 to be added to the total error score. On this basis, eight clear cases of error were selected for illustration and were scored as shown in Table 1, where the columns are headed 'H' for the human transcription using I.P.A. symbols, 'CME' for the "correct machine equivalent" (i.e. what the machine should have produced), 'WM' for the wrong machine transcription and 'S' for the error score for that segment.

TABLE 1
Examples of Error Scores

H	CME	WM	S
l	D	Fw	.6
d	S	Fm	.6
j	D	Fm	.8
d	S	D	.2
ʰ	B	Fs	.4
g	Fw	D	.6
Sil	Sil	S	.4
n	N	D	.6

RESULTS

Space does not allow a full presentation of the analysis of the example passages, but we will discuss one section: the first part is "Hello, operator" /ˈheləʊ ɒpəreɪtə /, which in equivalent machine symbols is FVDVVSBVDVSBV . Table 2 shows the H transcription ('H'), converted machine equivalent symbols ('CME'), durations (D1), the actual machine-transcribed symbols ('M') and their durations (D2). The right-hand column gives our evaluation. The error score for the extract chosen is calculated as 4.2, with 9.8 correct symbols, giving a success rate of 57%. The time-alignment score is calculated

TABLE 2

Sample Assessment of Errors

H	CME	D1	M	D2	Result	ErrorScore
h	F	7	Fw	7	correct	
e	V	10	V	18	correct	
l	D	9	Fw	14	wrong	.6
əʊ	V	37	V	16	correct	
			D	12	spurious	1
ɒ	V	24	V	16	correct	
p	S	5	S	9	correct	
ʰ	B	2	B	2	correct	
ə	V	6	V	33	correct	
ɹ	D	7	-	-	missed	1
eɪ	V	16	(V)	(")	(continuation)	
t	S	6	Fm	6	wrong	.6
ʰ	B	3	-	-	missed	1
ə	V	28	V	31	correct	

as 59%.

Overall scores for the whole of the chosen test material were calculated on the same basis: 92 segments were processed, with a success rate of 60%. On time-alignment, a total of 954 csec of speech was processed, with a success rate of 72%.

It is clear from the figures that our automatic segment marking is stricter than our previous technique: this is probably not a serious matter, since our chief concern is to have a technique that is reliable and objective, and which allows us to make comparative judgments about system performance under different conditions. More work to refine the technique is, however, still needed.

REFERENCES

[1] D.A.Klatt, 'Overview of the ARPA speech understanding project', in W.A.Lea (ed.) Trends in Speech Recognition, Prentice Hall, 1980.

[2] D.W.Shipman and V.W.Zue, 'Properties of large lexicons', IEE-ICASSP, 1982, pp.546-549.

[3] R.A.Cole, R.M.Stern and M.J.Lasry, 'Performing fine phonetic distinctions: templates vs features', in J.S.Perkell and D.A.Klatt (eds.) Invariance and Variability in Speech Processes, Erlbaum, 1986.

[4] J.Vaissière, 'Speech recognition: a tutorial', in F.Fallside and W.A.Woods (eds.) Computer Speech Processing, Prentice Hall, 1985.

[5] W.Jassem and P. Domagala, 'Phonetic segmentation in a bottom-up automatic speech analysis',in Proceedings of the International Conference on Speech Input/Output, Institute of Electrical Engineers, 1986.

[6] J.Dalby, J.Laver and S.M.Hiller, 'Mid-class phonetic analysis for a continuous speech recognition system', in Proceedings of the Institute of Acoustics, 8.7, pp.347-354, 1986.

[7] H.N.Roach and P.J.Roach, 'Automatic identification of speech sounds from different languages', Working Papers in Linguistics & Phonetics, University of Leeds, 1983.

[8] P.J.Roach, H.N.Roach and A.M.Dew, 'Assessing accuracy in automatic identification of phonetic segments', in Proceedings of the International Conference on Speech Input/Output, Institute of Electrical Engineers, 1986.

[9] J.C.Wells, 'A standardized machine-readable phonetic notation', in Proceedings of the International Conference on Speech Input/Output, Institute of Electrical Engineers, 1986.

[10] P.Ladefoged, A Course in Phonetics,(2nd ed.), Harcourt Brace Jovanovich, 1982.

[11] P.J.Roach, 'Rethinking phonetic taxonomy', in Working Papers in Linguistics & Phonetics, vol.4, 1986, Leeds University (to appear in Transactions of the Philological Society,1987).

[12] G.A.Miller and P.E.Nicely, 'An analysis of perceptual confusions among some English consonants', J.Ac.S.,27.2, pp.338-352, 1955.