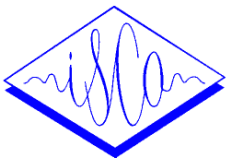
ISCA Archive
http://www.isca-speech.org/archive



2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

VERIFICATION OF LANGUAGE SPECIFIC PERFORMANCE FACTORS FROM RECOGNISER TESTING ON EUROM.1 CVC MATERIAL

Boerge Lindberg

Speech Technology Centre, University of Aalborg
Fredrik Bajers Vej 7, DK-9220 Aalborg, Denmark

10.21437/ICSLP.1992-386

ABSTRACT

This paper reports on the investigation of diagnostic and predictive assessment techniques in testing a recogniser on the EUROM.1 speech database, produced within the ESPRIT Project 2589 SAM.

Based on test results obtained on the full set of CVC-type (Consonant-Vowel-Consonant) words contained within the Danish Few Talker Part of the database, three different assessment techniques are investigated in this paper.

In the first, the substitution errors encountered are related to an analysis of the Danish phoneme inventory in terms of a representation of the phonetic distinctive features and the results show that many frequent substitutions found in a confusion matrix may be predicted as seen from the distinctive-feature representation.

In the second, a statistical analysis is carried out in which the asymmetric confusion matrices are transformed into symmetrical similarity matrices and applied in an individual differences multi-dimensional scaling analysis. This data-driven characterisation of the phoneme-inventory is illustrated in two-dimensional plots, and a high degree of correspondence with the distinctive feature map of phonemes is observed. This indicates possibilities of calibrating the performance of a recogniser against a given language by specifying the limitations of the recogniser in terms of phonetic dimensions specific to the language.

In the third, the estimated phoneme distances from testing a target recogniser on the EUROM.1 CVC material are used in a recogniser response model in order to aim at predicting the performance of the target recogniser on a different vocabulary. A limited experiment concludes that the observed performance of the target recogniser on the different vocabulary does not vary significantly from the predicted.

1. INTRODUCTION

Within the ESPRIT Project SAM, CVC-type words have been chosen as a key-corpus in the development of new methods for assessing recogniser devices and recognition algorithms. CVC-type words (i.e. words in which only a single phoneme varies) have also been extensively applied for intelligibility evaluation [1], and are attractive since they aim at representing the minimum phonetic differences occurring within a given language and since they contain all the potential confusions occurring on a minimum difference basis (open response set).

By applying a CVC-database, the aim is thus to outline diagnostic assessment procedures which give a more complete picture of the performance of the recogniser to be assessed, than otherwise obtainable by traditional assessment methods which most frequently are tied to a specific database and a corresponding application. A further discussion on alternative assessment techniques and associated principles for speech database selection is given in [2].

By means of a CVC-database it is desirable both to identify the limitations of the recogniser within the language and to predict the performance of that particular recogniser on a new vocabulary. As it is well known that many other factors are affecting the performance of a recogniser (such as environmental conditions), the present work should be seen as part of the initiatives within the ESPRIT SAM project to quantify these factors and to make qualitative descriptions of their effects on the performance.

The structure of this paper reflects the two purposes of the CVC-database. In section 2 a recogniser test experiment using EUROM.1 CVC-material is described. In the next two sections, the observed substitutions, given in the test results, are related to a priori phonological language knowledge and furthermore subjected to a multi-dimensional scaling technique, providing diagnostic information about the recogniser performance. Finally, in section 5, it is described how this information is applied in a recogniser response model in order to aim at predicting the performance on a different vocabulary.

2. RECOGNISER TESTING ON EUROM.1

The full EUROM.1 database, recorded for eight European languages, contains for each language both a Many Talker part and a Few Talker part. The 10 Few Talkers (five male and five female) were selected from the 60 Many Talkers in order to aim at representative recordings from a limited number of talkers, see [3] for further details about talker selection criteria and the full EUROM.1 database.

Table I

Contents and phonemic transcription of the Danish CVC-lists on EUROM.1 (_ (underscore) denotes the variable phoneme).

Initial Consonants	_al@, _il@, _ul@
Var. phonemes (SAMPA):	p t k b d g f v s h m n l j R
Intial Consonant Clusters	_RAl@, _jal@
Var. phonemes (SAMPA):	p t k b d g
Intial Consonant Clusters	_lal@
Var. phonemes (SAMPA):	p k b g
Intial Consonant Clusters	_val@
Var. phonemes (SAMPA):	t k d g
Final Consonants	la_
Var. phonemes (SAMPA):	d D s l n
Final Consonants	lA_
Var. phonemes (SAMPA):	b g f m N
Vowels	t_:d@, t_d@
Var. phonemes (SAMPA):	i e E { A y 2 u o O Q
Diphthongs	t_id@
Var. phonemes (SAMPA):	Q A
Diphthongs	t_ud@
Var. phonemes (SAMPA):	y 2 i e E A O Q
Diphthongs	t_Qd@
Var. phonemes (SAMPA):	a i e y 2 u o

For Danish, the CVC-lists contain such sequences as CVC's, CCVCV's, CVCV's and CVVCV'c as well as different manifestations of short and long vowels and are structured as shown in the above Table I (phonemes are shown in SAM Phonetic Alphabet (SAMPA)). From Table I it can be seen that the Danish CVC speech database comprises 114 words distributed on 16 different CVC-word lists with varying positioning (initial, central or final) of the single phoneme being variable within the list.

In the work reported here, a speaker dependent isolated word DTW recogniser is applied as the target recogniser.

The assessment procedure is conducted utilising the SAM reference standard work station, SESAM, and using SAM software packages developed for assessment control and scoring of results [4]. The target recogniser is executing on a PC Board and thus enables a standard SAM test setup in which the recogniser is located inside the SESAM work station, thereby enabling a compact hardware test setup.

2.1 Training and Testing

All the CVC lists are uttered five times by the 10 Few Talkers (only the Few Talkers made CVC-recordings on EUROM.1). These list-recordings are applied in a leave-one-out test of the recogniser, so that one list-recording is used for training and the other four list-recordings are used for testing. This procedure is then repeated five times for each talker, giving a total of 50 test sessions for each CVC list.

This test sequence does not take into account the possibility of having substitutions between the CVC lists, i.e. it does not allow for substitution possibilities between e.g. short and long vowels.

2.2 Results

Based on the above leave-one-out training and testing methodology the following confusion matrices were obtained by using the SAM scoring software package, SAM_SCOR [5] (for reasons of space these are only showed for the three initial consonant lists and the two vowel lists). The matrices shown in Table II and III below are the accumulated confusion matrices in the sense they present the overall sum of confusions in the CVC-word lists involved. Thus the number of confusions between <p> and <k> represent the sum of confusions between (pal@, kal@), (pil@, kil@) and (pul@, kul@).

The tables below show approx. 80% correct recognition rate on vowels and approx. 70% correct recognition rate on initial consonants.

The matrices below are then subjected to three different analyses, reported in the following sections.

Table II

Accumulated confusion matrix, vowels in contexts t_d@ and t_:d@ (short and long vowels). Stimuli are shown horizontally, responses vertically.

	i	e	E	{	A	y	2	u	o	O	Q
i	308	14	3	.	1	12	.	2	1	2	1
e	42	283	78	4	.	10	5	.	.	.	.
E	8	85	256	59	3	6	9	.	.	.	.
{	1	8	49	315	1	8	5	5	1	2	3
A	2	.	.	.	374	.	.	.	1	1	14
y	1	1	.	.	1	308	24	1	1	1	.
2	2	.	.	4	.	30	341	4	.	2	1
u	.	.	.	.	.	1	2	317	40	5	1
o	.	.	.	.	.	2	1	52	290	28	11
O	.	1	.	.	1	1	2	18	55	317	44
Q	.	.	.	1	10	1	1	.	4	32	321

Table III

Accumulated Confusion Matrix, Initial Consonants, varied in three different contexts, _al@, _ul@ and _il@. Stimuli are shown horizontally, responses vertically.

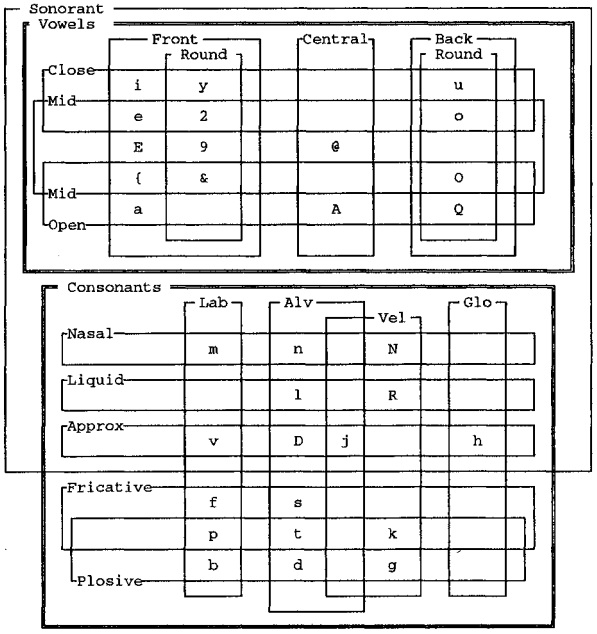
	p	k	t	b	g	d	f	v	s	h	m	n	l	j	R
p	424	42	35	8	8	3	44	.	1	80	.	.	.	.	.
k	19	447	7	.	6	.	5	.	6	45	.	.	.	.	2
t	15	20	487	.	1	3	27	.	87	12	.	.	.	.	.
b	13	1	4	360	36	125	73	59	22	18	13	3	6	5	22
g	16	17	4	24	444	31	16	5	8	25	6	1	4	7	9
d	9	2	2	95	56	375	29	16	34	15	3	3	8	11	4
f	13	1	15	38	10	16	291	18	54	18	1	.	1	1	6
v	3	.	.	36	2	16	20	402	7	6	60	19	30	1	19
s	2	2	23	13	3	15	46	4	341	13	.	.	8	.	2
h	86	68	23	9	20	9	40	5	21	352	.	2	11	5	4
m	.	.	.	2	.	1	.	40	.	1	408	78	1	1	5
n	.	.	.	.	.	.	.	13	.	.	96	461	15	8	5
l	.	.	.	1	.	2	.	15	1	.	3	21	489	17	7
j	.	.	.	1	5	.	.	1	2	9	.	6	27	544	.
R	.	.	.	13	9	4	5	22	3	2	10	6	.	.	511

3. DISTINCTIVE FEATURE DESCRIPTION

In order to relate the observed phoneme confusions to a priori phonological language knowledge, the individual phonemes are described by a set of distinctive features indicating place and manner of articulation. E.g. the Danish phoneme /e/ (using SAMPA symbols) is a sonorant front unrounded and mid-close vowel, whereas /p/ is a non-sonorant an unvoiced, labial and plosive consonant. This is illustrated in Table IV, in which the Danish phonemes are arranged according to their place and manner of articulation, using the SAMPA symbols [6].

Table IV

Danish Distinctive Vocalic and Consonantal Features



In relating the confusions observed in Table II to the above phoneme inventory as given in Table IV, it can be seen that substitutions between the front and back vowel groups almost never occur. In both of these groups the results indicate furthermore that substitutions can be observed between neighboring phonemes such as open/mid-open, mid-open/mid, mid/mid-close and close/mid-close vowels. However, almost no substitutions are identified between open and close vowels.

The substitutions in the consonantal CVC-word lists (Table III) are not as clearly identifiable as those in the vowel lists.

However, the results have shown a clear confusion within the group of plosive fricative consonantal phonemes /p, t, k/, and within the group of plosive consonantal phonemes /b, d, g/. Some confusions are also identified across the two groups, but these are primarily found in either alveolar or velar phonemes.

The above analysis indicates that language specific factors affecting the performance of a given recogniser can be identified to some extent from an analysis of the phoneme-inventory of the application vocabularies in terms of distinctive-feature representations.

However, the high number of confusions between /p/ (labial, plosive fricative consonant) and /h/ (glottal, approximant consonant) identify limitations to the approach, as they are probably caused by the fact that these phonemes apparently exhibit similarities in their acoustic realisation due to the circumstance that many different articulatory configurations may perceptually be the same sound.

4. MULTIDIMENSIONAL SCALING

Based on the asymmetric confusion matrices shown in tables II and III in section 3, symmetric similarity matrices can be generated according to the procedure presented in [7]. In the following, these individual phoneme similarities are subjected to a multi-dimensional scaling technique [8,9], in which the best positioning of the individual points (phonemes) is estimated. The results of this analysis, restricted to two dimensions, are shown below in Fig. 1 and 2 for the vowels and initial consonants, respectively.

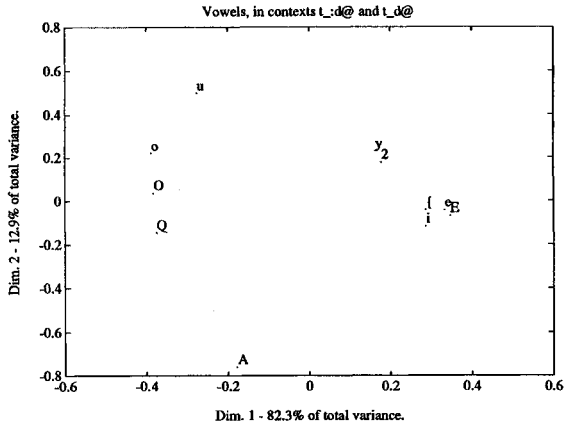


Figure 1 Two-dimensional scaling plot for the vowels.

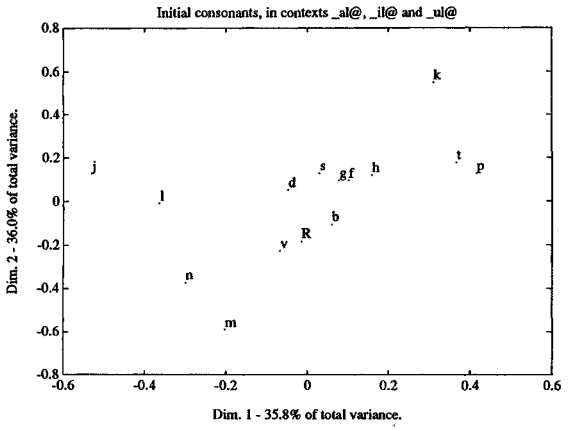


Figure 2 Two-dimensional scaling plot for initial consonants.

By visual inspection of the above Fig. 1 and 2, a high correspondence is observed with the distinctive feature map given in Table IV. This is most significant in the vowel space, in which case the inter-group positioning of the group of back-rounded vowels /u, o, O, Q/, the group of front-rounded vowels /y, 2/, the group of front-unrounded vowels /i, e, E, {/ and the open central phoneme /A/ are clearly distinct and fully in accordance with the distinctive feature map. This is also the case with the intra-group positioning of the phonemes within the back-rounded vowels, whereas the group of front-rounded vowels are not as clearly separated from each other. It should be observed from Fig. 1 that approx. 95% of the total variance is described by means of a two-dimensional scaling, which may indicate that the above data-driven representation is a good characterisation of the vowel-space in general.

In the consonantal space the correspondence is not so apparent although the main classes of phonemes are grouped in accordance with the distinctive feature map. However, a two-dimensional plot seems in this case not to be a fully sufficient characterisation of the initial consonants space, as only 71.8% of the total variance is covered. This insufficient coverage of variance may explain some of the mismatching between the distinctive feature map in Table IV, and the consonantal map in Fig. 2, as some of the phoneme distances observed in Fig. 2 clearly do not correspond to the originating confusion matrices, see Table III. Thus for what concerns the consonantal space, it seems required to apply more than two dimensions in the multi-dimensional scaling procedure.

Based on the above considerations it may however be concluded that the multidimensional scaling analyses may lead to a procedure in which the limitations and performance of a recogniser may be calibrated against the acoustic-phonetic dimensions specific for the language.

5. VOCABULARY DIFFICULTY PREDICTION

In the following experiment the aim is to apply the results obtained on recogniser testing on EUROM.1 data to predict potential confusions on an alternative vocabulary, which is different from EUROM.1 CVC's. This is done by applying the HENR-model [10,11] in which the inter-phoneme distances are obtained from the multi-dimensional scaling procedure described in section 5. In this case the dimension was set to four, which is the maximum available within the INDSCAL program [9], applied for the multi-dimensional scaling.

The Danish digits were chosen as the new target vocabulary, and a predicted confusion matrix was obtained, based on the phonotypical transcription of the digits in Danish, see Table V below. As the reported experiment regarding predictive assessment was very limited, it was not possible to calibrate the HENR-model against the originating performance figures from EUROM.1. Thus, in order to be able to compare the predictions with the observations mentioned below, it was necessary, as input to the HENR-model, to use the observed overall performance of 94.8% from Table VI, as a basis for the prediction of potential confusions.

In order to aim at verifying the predictions, the target recogniser was tested on the Danish digits on the SAM EUROM.0 speech database. For the four talkers on EUROM.0, two series of tests were conducted in which 1 repetition was used for training and 10 for testing (out of the 20 available pronunciations of the 10 digits). The resulting confusion matrix is shown below in Table VI.

Table V

Predicted confusion matrix for the target recogniser on Danish digits.

CONFUSION MATRIX BASED ON AN OVERALL PERFORMANCE OF 94.8 %

	en	to	tRE	fi:Q	fEm	sEgs	syu	O:dQ	ni	nOl
en	91.5	.	0.7	.	7.1	0.6	.	.	.	.
to	.	96.7	2.0	.	.	0.4	0.6	.	.	.
tRE	0.7	1.8	86.2	.	1.8	7.7	.	.	1.7	.
fi:Q	.	.	.	99.9	.	.	.	.	.	.
fEm	6.9	.	1.8	.	88.8	2.4	.	.	.	.
sEgs	0.6	0.4	7.9	.	2.4	88.6	.	.	.	.
syu	.	0.6	.	.	.	.	99.3	.	.	.
O:dQ	.	.	.	.	.	.	.	99.9	.	.
ni	.	.	1.9	.	.	.	.	.	97.6	0.2
nOl	.	.	.	.	.	.	.	.	0.2	99.5

Table VI

Actually observed confusion matrix from testing the target recogniser on the Danish digits on the EUROM.0 speech database.

	en	to	tRE	fi:Q	fEm	sEgs	syu	O:dQ	ni	nOl
en	80	.	.	5	.	.	.	.	.	.
to	.	79	.	.	.	.	1	.	.	.
tRE	.	.	72	.	.	10	.	.	.	.
fi:Q	.	.	.	75	.	.	.	.	.	.
fEm	.	.	3	.	80	11	3	.	.	2
sEgs	.	.	5	.	.	59	.	.	.	.
syu	.	.	.	.	.	.	75	.	.	.
O:dQ	.	.	.	.	.	.	.	80	.	.
ni	.	.	.	.	.	.	.	.	80	.
nOl	.	.	.	.	.	.	.	.	.	78

Although this was a very limited experiment, the preliminary conclusion may be drawn that, if the recogniser performance is first measured on the CVC-sets of the EUROM.1 database, then the derived estimated inter-phoneme distances together with the HENR-model may be a valuable mean of predicting the performance of a given recogniser on a new vocabulary, as most of the major predicted confusions were actually observed when conducting the recogniser testing.

6. CONCLUSION

This paper has reported on the application of the Danish Few Talker Part of the EUROM.1 database in recogniser testing and diagnostic as well as predictive assessment techniques have been presented.

The confusions observed in the testing were related to an analysis of the Danish phoneme inventory in terms of a phonetic distinctive feature representation and the results showed that most of the frequent substitutions observed from a confusion matrix were predictable as seen from the distinctive-feature representation. Thus these features may to some extent be the basis for a performance calibration relative to the language applied, although the acoustic similarities exhibited by some phoneme pairs need to be further modelled.

The observed confusions were subjected to a multi-dimensional scaling algorithm from which two-dimensional plots of estimated phoneme positions were generated. These plots in the two-dimensional sub-space showed an apparent correspondance with the relative positioning of the vowels within the distinctive feature map representation, whereas the correspondence in the consonantal space was restricted to subgroups of consonants.

The estimated inter-phoneme distances were applied as input to the HENR recogniser response model. This model was used to predict the performance and potential confusions within a new vocabulary different from EUROM.1 CVC-words. These predictive performance estimates were then compared to performance figures actually observed when testing the recogniser on that particular new vocabulary. A preliminary conclusion from this experiment is that if recogniser performance is first measured on the CVC-sets of the EUROM.1 database, then the derived estimated inter-phoneme distances together with the HENR-model may be a valuable mean of predicting the performance of a given recogniser on a new vocabulary.

However, this latter conclusion needs to be manifested by further experiments, even in the limited isolated word domain that formed the basis for the experiments reported.

Acknowledgments

The research reported has partly been funded by the Danish Technical Research Council by the frame programme "Spoken Language Processing in Application Oriented Dialogue Systems" and has partly been done within the ESPRIT Project 2589 SAM.

In preparing this paper, the author would like to especially acknowledge the assistance received from Sven Danielsen, Research and Development Dept., Jydsk Telefon, Denmark as well as Dr. Roger K. Moore, DRA-SRU, Malvern, England and Prof. Paul Dalsgaard, Speech Technology Centre, University of Aalborg, Denmark for their valuable comments to the work presented.

References

[1] H.J.M. Steeneken (1987): "Diagnostic Information from Subjective and Objective Intelligibility Tests", Proc. ICASSP, Dallas.

[2] H.J.M. Steeneken (1991): "RAMOS - Recognizer Assessment by Means of Manipulation of Speech Applied to Connected Speech Recognition", Proc. EUROSPEECH 91, Genova, Italy

[3] T. Sherwood (1992): "Guide to EUROM.1 Speech Database", SAM-NPL-102, ESPRIT 2589 Project SAM General Publication, March 1992.

[4] ESPRIT Project 2589 SAM (1992): "Final Report - Year Three", SAM-UCL-G004, University College London, England.

[5] B. Lindberg, R. Joergensen, O. Andersen, S. Danielsen (1992): "SAM_SCOR v. 3.1 Reference Guide", Document SAM-IES-66, March 1992, ESPRIT Project 2589 SAM.

[6] P. Dalsgaard, O. Andersen, W. Barry (1991): "Multi-Lingual Acoustic-Phonetic Features for a Number of European Languages", Int. Conf. EUROSPEECH, Sep. 1991, Genova.

[7] A.J.Bosman (1989): "Speech Perception by the Hearing Impaired", Doctorial dissertation, Rijks-Universiteit, Utrecht

[8] J.D. Carroll, J.-J. Chang (1970): "Analysis Of Individual Differences in Multidimensional Scaling via an N-Way Generalization of "Eckart-Young" Decomposition", Psychometrika - Vol. 35, No. 3, Sep. 1970

[9] J.B.J. Riemersma (1974) : "Development of an individual differences multidimensional scaling program", Internal document, TNO Soesterberg, Netherlands.

[10] R.K. Moore (1977): "Evaluating Speech Recognisers", IEEE Trans. on ASSP, Vol. ASSP-25, No.2, April 1977

[11] R.N. Shepard (1957): "Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in a Psychological Space", Psychometrica, vol. 22, pp. 325-345, Dec. 1957