HUMAN AND MACHINE IDENTIFICATION OF CONSONANTAL PLACE OF ARTICULATION FROM VOCALIC TRANSITION SEGMENTS

Andrew C Morris, Gerrit Bloothooft**, William J Barry***, Bistra Andreeva***, Jacques Koreman****

* *Speech and Hearing Research Group, Sheffield University, UK, Tel. +44 (0)114 222 1907, E-mail: a.morris@dcs.shef.ac.uk*

** *Utrecht Institute of Linguistics OTS, Holland, Tel. +31.30.2536042, E-mail: gerrit.bloothooft@let.ruu.nl*

*** *Institute of Phonetics, University of Saarbrücken, Germany, Tel. +49(0)681.3024500, E-mail: wbarry@coli.uni-sb.de*

ABSTRACT

One of the most difficult problems in the first stages of automatic speech recognition (ASR) is the identification of consonantal place of articulation (CPA). It is known that the acoustic correlates for CPA reside largely in the pattern of formant transitions preceding vocal tract closure and following release, but common speech preprocessing techniques make only a limited attempt to capture these spectral dynamics in the representation which they pass on for recognition. In order to test alternative preprocessing strategies, we have prepared a multilingual set of VC and CV vocalic transition segments and then compared the baseline performance of human perception of CPA in this dataset with the performance of two common ASR techniques. Representaions initially tested were concatenated mel cepstra and mel ceptra plus cepstral differences.

Keywords formant transitions, place perception, Kohonen map, Gaussian mixture classifier

1. INTRODUCTION

1.1 The importance of spectral dynamics

It is well established that the two main sources of spectro-temporal information governing the perception of CPA [4,6,8,9] are:

- the characteristics of the static release or closure spectrum (to be referred to as the RC spectrum)
- the formant transitions over a period of order 50ms preceding consonantal closure and following consonantal release [4].

While the greatest concentration of CPA information is generally in the release or closure (RC) spectrum, the relative weight of these two sources of information depend on vowel context and noise conditions.

The value of formant transition patterns is that they reflect the articulatory gestures towards or away from the target consonant, even when the consonant itself is not fully realised [2,3]. High performance in ASR is therefore dependent on the effective exploitation of these patterns across time as well as across frequency. However, the techniques commonly used for speech coding prior to recognition, such as cepstral coefficients, plus first and possibly second order cepstral differences, are able to capture only a part of these dynamics.

By comparing the performance of human perception with a range of different ASR techniques in identifying a multilingual set of VC and CV vocalic transition segments, weshould be able to identify the principle areas in which present speech preprocessing techniques are deficient, and possibly draw some conclusions about how the coding of spectral dynamics can be improved.

1.2 Separating dynamic cues from static

We aim here to examine the role of formant transitions preceding consonantal closure and following consonantal release in determining CPA, in a multilingual set of purely vocalic VC and CV transition segments. In work such as [3] the whole interval around the transition centre was used in perception tests, and elsewhere [6,7] for machine recognition. In separating the vocalic transition region from the consonantal release or closure, we are focusing on just one of these two important sources of information.

By separating the vocalic region on one side of the RC spectrum from the voicing or frication on the other side of this spectrum, as well as the RC spectrum itself, we have also systematically removed voice onset time (VOT) information, another strong cue for CPA. With RC spectra *and* VOT removed it is not surprising that these tests *were not easy for either humans or machines*.

2. EXPERIMENTAL DESIGN

2.1 Database preparation

The data for our CPA identification tests consists of 35 ms voiced speech segments taken from immediately before consonantal closure and after consonantal release, i.e. from the vocalic portions of the transitions only. The

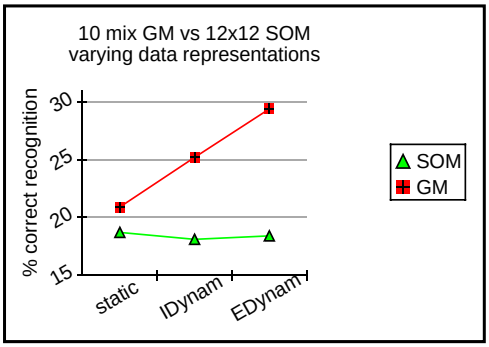

Fig.1

Fig.1 *shows 10 mix Gaussian classifier ASR recognition score for pattern vectors constructed as static: (f1+f2+f3)/3, implicit dynamic: (f1,f2,f3) and explicit dynamic: (f1, f2, f3,f2-f1,f3-f2).*

Fig.2 *shows the effect of varying the number of Gaussian mixture components for the best scoring representation.*

Fig.3 *shows the effect on GM of refining the parameters obtained by data clustering by a number of EM iterations.*

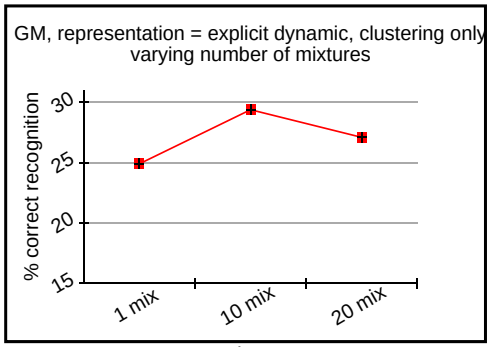

Fig.2

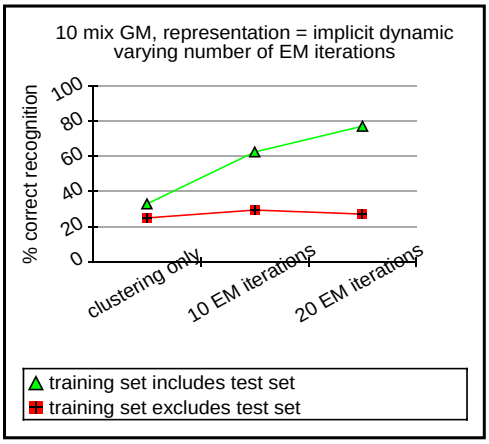

Fig.3

ba	da	ga	pa	ta	ka
bi	di	gi	pi	ti	ki
bu	du	gu	pu	tu	ku

Fig.4
Button response menu for place-manner perception test

Table 1.

Phoneme place-manner group labels

	voiced	unvoiced
labial	b = bmv	p = pf
alveolar	d = dnzl	t = tTs
velar	g = gNrR	k = kx
open	a = aA@V	
front	i = ieIE	
back	u = uUoOQ	

full data set was made up from all combinations of the following attributes from the hand labelled EUROM.0 "number passage" corpus, giving 9000 segments :

- 5 languages (English, Danish, Dutch, German, Italian)
- 5 speakers
- 5 examples
- 2 sexes
- 2 contexts (VC, CV)
- 3 vowel places of articulation
- 2 consonant manners (voiced, unvoiced)
- 3 consonant places of articulation

The phoneme labelling used is close to SAMPA (Speech Assessment Methodology Phonetic Alphabet), with place-manner group labels as shown in Table 1.

2.2 Human CPA perception tests

The human perception test was taken by 10 male and 10 female listeners from each of two nationalities: German and Dutch. Each listener was tested on VCs and CVs separately, on either male or female speech. Each test set comprised one example of each transition pair place-manner category from each male or female speaker, making up 225 examples, with the response menu appearing as in Fig.4.

The test set was presented in a different random order for each subject, with each 35ms stimulus played twice in short succession, then repeating every three seconds, until a response was recorded. In this paper we present only a summary of the full analysis of these perception test results which will be published else where for their interest to phonetics. Results in Table 2 are for German male and female subjects combined.

2.3 Machine CPA recognition tests

Every transition in the data set, minus the perception test set, was used for ASR training. The human perception test set was used for testing. The automatic recognition systems tested were the Kohonen Self Organising Map (SOM) [1] and the Gaussian mixture classifier (GM) [1]. Each 35 ms speech segment was parametrised as 3 frames of 12 Mel Frequency Cepstral Coefficients (MFCC) (zeroth or energy coefficient excluded). Frames were 15 ms at 10 ms centres.

2.3.1 Data representations tested

The three data representations tested were formed from three consecutive MFCC frames (f1,f2,f3) as follows:

Static	(f1 + f2 + f3)/3
Implicit dynamic (ID)	(f1, f2, f3)
Explicit dynamic (ED)	(f1, f2, f3, f2-f1, f3-f2)

Test results pooled for 20 German subjects

	bV	dV	gV	tot
	average 41.0% correct			
bV	51.4	22.2	26.4	1500
dV	35.4	38.2	26.4	1500
gV	40.2	26.4	33.4	1500
tot	1905	1302	1293	4500

	Vb	Vd	Vg	tot
	average 47.9% correct			
Vb	52.9	19.9	27.3	1390
Vd	30.5	44.4	25.1	1500
Vg	30.0	23.4	46.6	1400
tot	1612	1270	1408	4290

Test results for statistical (10 Gaussian Mixture) classifier

	bV	dV	gV	tot
	av. 46.2% correct			
bV	43.3	41.3	15.3	150
dV	12.7	80.7	6.7	150
gV	25.3	60.0	14.7	150
tot	122	273	55	450

	Vb	Vd	Vg	tot
	average 48.1% correct			
Vb	42.1	48.6	9.3	140
Vd	10.7	74.7	14.7	150
Vg	17.9	56.4	25.7	140
tot	100	259	71	430

Table 2

Table 2 shows consonant place confusion matrices, in CV and VC context, for German listeners. Percentage confusion scores are shown for human perception (top) and GM classifier (bottom). Voiced and unvoiced place groups are pooled and labelled b, d, g, while all vowels are grouped into one group, labelled V. Row is true class, and column is class identified. GM classification shows very strong "d" dominance, whereas human perception shows a slight "b" dominance, with errors more evenly distributed between classes.

2.3.2 Kohonen map tests

Various SOM configurations were tested. Best results were obtained with a 12 x 12 grid, after 500 training iterations through the full training set (epochs). Update radius and learning rate used exponential decay, with half lives of proportion 0.3 and 0.3 of the fixed total number of iterations, and initial radius and learning rate of 12 and 0.2. Performance was by optimised using a majority vote from the top k = 2 response classes, having tested k from 1 to 10.

2.3.3 Gaussian classifier tests

The GM classifier was tested with between 1 and 20 mix components. Both full and diagonal covariance matrices were tested. Although pattern vector components were correlated, diagonal covariance was found to give significantly better results. Parameter estimation used an initial clustering algorithm followed by an Expectation Maximisation (EM) based iterative refinement procedure [1]. EM iteration is only theoretically guaranteed to converge to a local optimum solution, so it was important that the clustering algorithm used should obtain an initial estimate close to the global optimum.

The clustering used was a splitting algorithm in which, starting with all data in one cluster, in each splitting stage a new cluster is assembled for which the ratio of the resulting total between cluster square Euclidean distance to the total within cluster square distance is maximum.

EM iteration did not significantly improve performance here unless the training set included the test data. This suggests that for the limited amount of training data available, EM iteration only resulted in overfitting.

3. RESULTS

Table 2 shows CV and VC CPA confusion tables for the perception test and for the best GM classifier.

The GM classifier gave best results with the Explicit Dynamic representation [Fig.1] - which is similar to that used in existing high performance CDHMM based ASR systems. Ten mixes [Fig.2], with diagonal covariance, achieved the best compromise between model flexibility and overfitting [Figs.2,3].

The SOM gave best results with the Static representation, but these were well below GM classifier performance. It appears that the SOM is not able to take advantage of difference coefficients, even though these are known to carry useful information. This may be due to the Euclidean distance measure being less than optimal when zero and higher order differences share the same data space. This problem of relative scaling does

not arise in the case of diagonal covariance GM, where each component is treated independently.

Perception test results were very similar for all combinations of speaker and listener nationality. ASR test results were also similar for different speaker nationalities. However, ASR results were consistently different from perception results, and in a similar way for each of the five languages tested.

4. DISCUSSION

While the CPA error rate for both humans and machines is very high (which not unexpected after the exclusion of both RC spectra and VOT), humans always identify the correct place more than any other. This not the case with the best machine arrangement, for which recognition of velar place is very weak, particularly in CV context. The cues used in human place recognition are therefore different from those available to our best ASR system.

The GM classifier with a large number of mixes can model very closely the true data distribution and should therefore approach *optimal* performance, limited only by the amount of training data available. If we accept that the classifier is near optimal for the training data available, and that the quantity of training data is reasonable, then the deficiencies in the present GM based classifier are at least in part due to suboptimal speech data coding. From Fig.1 it appears that ASR recognition performance increases steadily as the coding of spectral dynamics becomes progressively more direct.

5. CONCLUSION

The ED pattern vector codes both spectral energy and energy change in time directly. What appears to be missing from this representation is any direct measure of energy change across frequency, and of energy change across frequency and time together, or spectral slope.

As a first attempt at coding directly for spectral slope, we could investigate appending the following "diagonal difference" coefficients to the frame coefficient x(f,t):

$$x(f+1,t+1) - x(f,t)$$
$$x(f,t+1) - x(f+1,t)$$

As well, or alternatively, we could add non linear terms, such as the following "cross products":

$$x(f,t)*x(f+1,t+1) - x(f+1,t)*x(f,t+1)$$

Given a sequence of consecutive cepstral, spectral, PLP or other data frames, $F = (f_1,f_2,...f_n)$, besides making explicit certain data features, almost any non linear projection g(F) of this data will tend to have some beneficial effect on recognition, because of its tendency to enhance both SNR and data orthogonality [5].

We should also note that the projection used at any time t may benefit by varying in accordance with certain gross features as detected by some function f(F), such as stationarity [6,7] or texture quality.

In seeking a way forwards from here, we should consider looking to phonetics [2] to indicate the invariants which we should be trying to make explicit, and to auditory physiology [6] to get some idea of the kind of feature detectors used by the auditory system.

ACKNOWLEDGEMENTS

This work was supported by the EU HCM Network contract CHRX-CT93-0098.

REFERENCES

[1] Bishop, C. M. (1995) ***Neural networks for pattern recognition***, Clarendon Press.

[2] Fant, G. (1960) Acoustic theory of speech production, Mouton & Co., The Hague.

[3] Furui, S. (1986) "On the role of spectral transition for speech perception", JASA 80(4), pp.1016-1025

[4] Kewley-Port, D. (1986) "Converging approaches towards establishing invariant acoustic correlates of stop consonants", in ***Invariance and variability in speech investigation***, (J.S.Perkel & D.H.Klatt eds.), MIT Press, pp.193-197.

[5] Kohonen, T. (1987) ***Self-organising and associative memory***, 2nd edition, Springer series in information sciences Springer-Verlag.

[6] Morris, A. C., Schwartz, J.-L., & Escudier, P. (1993) "An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram", Computer Speech & Language, 2, pp.121-136.

[7] Morris, A. C. & Pardo, J. M. (1995) "Phoneme transition detection and broad classification using a simple model based on the function of onset detector cells found in the cochlear nucleus", Proc. Eurospeech'95, pp.115-118.

[8] Nathan, K. S. & Silverman, H. F. (1991) "Classification of unvoiced stops based on formant transitions prior to release", Proc. ICASSP'91, pp.445-448.

[9] Wang, W. S. J. (1959) "Transition and release as perceptual cues for final plosives", Speech and Hearing Research, 2, pp.66-73.