FROM: PROCEEDINGS OF THE 1985 IEEE INTĹ. CONF. ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, MARCH 1985, PP. 17-20.

EXPERIMENTS ON SPEAKER-INDEPENDENT RECOGNITION OF HAND-SEGMENTED FRENCH VOWELS

Matthew Lennig
Bell-Northern Research
3 Place du Commerce
Nuns' Island, Montreal, Quebec
Canada H3E 1H6

Abstract

This paper addresses the problem of recognizing hand-segmented vowel phonemes in isolated words pronounced by different speakers. Overcoming variation in the spectral properties of similar vowels produced by different speakers (vowel normalization) is a subproblem in the goal of large vocabulary, speaker-independent speech recognition. Two normalization techniques are compared in a series of recognition experiments on the 15 oral and nasal French vowel phonemes.

1.0 Introduction

Eskenazi and Liénard have shown that human listeners can phonemically classify isolated French vowels pronounced by ten speakers with 88 percent accuracy. However, their automatic classification algorithm, based on spectral similarity, achieves a recognition rate of only 50 percent.[1] The poor performance they observe in automatic vowel recognition is largely due to spectral variability in different speakers' voices. While the human listeners are able to ignore differences related to speaker type and focus on phonemic class, the spectral classification technique is not. The purpose of a *normalization* method is to transform spectral representations in such a way as to minimize speaker-related information while preserving phonemic information.

Two normalization techniques are compared in a series of automatic phoneme recognition experiments on French vowels.[2] The first normalization technique involves rotating the vowel space of each speaker in the multidimensional spectral space so that directions of maximum variance for all speakers coincide. This method, which has never been tested empirically, is equivalent to a *principal components* representation of each speaker's vowels in which the principal direction eigenvectors are determined individually for each speaker.[3] The second technique involves *centering* the vowel space by removing the overall speaker mean.

2.0 Experimental Methods

The data consist of a list of ninety French words, mainly monosyllables, read by ten speakers of Parisian French, five women and five men. In the case of polysyllabic words, only the final stressed syllable is used. Vowels are segmented by hand, yielding 891 vowel tokens (ninety per speaker, less misread and missing words). After preprocessing and optional normalization, each vowel token is represented by a single, D-dimensional parameter vector averaged over the duration of the vowel, excluding consonant transitions. Averaging is justified because Parisian French vowels tend to be steady state.

Speaker-independent recognition experiments, in which the test speaker is excluded from the training set, measure relative effectiveness of the normalization methods. Each speaker-independent recognition experiment consists of ten subexperiments in which each of the ten speakers serves as test speaker, the nine others being used for training (leave-one-out procedure). Error rates are pooled across the ten subexperiments.

For all experiments, recognition is accomplished using fifteen D-dimensional reference vectors, one corresponding to each phoneme averaged over the training set. Test tokens are recognized by calculating the Euclidean distance to each of the fifteen references and applying the nearest-neighbor decision rule. Two phonemic confusions are not counted as errors: /ɛ̃/ = /œ̃/ and /a/ = /ɑ/. The first of these pairs is completely merged in Parisian French; the second is merged for many but not all speakers.[4] Confusion matrices also show much confusion between the phonemes /ɔ/ and /œ/. This is because the allophone of /ɔ/ that occurs before consonants other than /ʀ/ is completely overlapped with /œ/.[5] The latter confusion is nonetheless included in the error scores, leading to a conservative estimate of recognition performance.

[1] M. Eskenazi and J.S. Liénard (1983), "Recognition of steady-state French sounds pronounced by several speakers: comparison of human performance and an automatic recognition algorithm", *Speech Communication* **2**, pp. 173-177. The one essential difference between Eskenazi and Liénard's database and the one to be described in this paper is that the former was produced by having trained speakers pronounce vowel phonemes in isolation, while the latter uses untrained subjects reading word lists.

[2] This work was funded by DCIEM, Department of National Defence, Canada. The author thanks Pascal Auxerre of the École Nationale Supérieure des Télécommunications (Paris) for his help in executing these experiments.

[3] M.M. Taylor (1973) argues that the nervous system may actually perform adaptive principal components analyses in "The problem of stimulus structure in the behavioural theory of perception", *South African Journal of Psychology* **3**, pp. 23-45.

[4] See M. Lennig (1978), *Acoustic Measurement of Linguistic Change: the Modern Paris Vowel System*, University of Pennsylvania Dissertation Series, No. 1, U.S. Regional Survey, 204 North 35th St., Philadelphia, PA 19104.

[5] *Idem.* Also, A. Martinet (1958), "C'est jeuli le Mareuc", *Romance Philology* **11**, pp. 345-355.

3.0 Speaker-Dependent Recognition

To establish an upper bound on speaker-independent recognition accuracy, speaker-dependent recognition experiments are performed first. Each speaker-dependent experiment similarly consists of ten subexperiments, except that the training set contains *only* the test speaker's own tokens, including the test token itself.

Four different parameter sets are compared: the first D of the 20 *log channel energies*, $(L_1, \ldots, L_D)$, outputs from a mel-frequency channel bank; the mel-based *cepstrum*, $(C_0, \ldots, C_{D-1})$, which is the cosine transform of the log channel energies;[6] the mel-based cepstrum, $(C_1, \ldots, C_D)$, excluding C_0, the overall energy component; and the *global principal components*, $(G_1, \ldots, G_D)$, based on the global covariance matrix of the log channel energies.[7]

Figure 1 shows percentages correct recognition in speaker-dependent experiments using the four parameter sets listed above as dimensionality D varies from 2 to 19. Recognition performance using principal components, cepstrum including C_0, and log channel energies converge at higher dimensions since they are simply rotations of each other. When fewer dimensions are used, principal components perform better than either the cepstrum including C_0 or the log channel energies since the first few principal components express most of the variance in the data.

Perhaps the most striking result appearing in Fig. 1 is the advantage to be gained by *excluding* the overall energy parameter, C_0. This implies that as far as vowel recognition is concerned, variance due to overall amplitude does not aid in vowel recognition and should be considered as noise. The mel-based cepstrum lends itself to the simple elimination of overall energy by omitting C_0. To eliminate overall energy from the principal components parameterization, however, requires an energy equalization prior to computation of the covariance matrix (see Sect. 4.1).

To compare the performance of principal component parameters based on individually determined covariance matrices as opposed to one global covariance matrix, a speaker-dependent experiment was conducted at $D = 8$. Use of *individually* determined principal component parameters, $(I_1, \ldots, I_8)$, resulted in a speaker-dependent recognition rate of 82 percent correct. This is slightly worse than the 85 percent achieved with global principal components and substantially worse than the 88 percent achieved with the mel-based cepstral coefficients, $(C_1, \ldots, C_8)$.

4.0 Speaker-Normalization Experiments

Normalization effectiveness is measured two ways: recognition accuracy, described above, and a statistical measure of separation, F, based on Fisher's F-ratio. F is the ratio of between-phoneme variance to within phoneme variance, where variance is interpreted as the average squared distance to the appropriate mean vector. Larger values of F indicate more separation between phonemes and thus better normalization methods.

4.1 Principal Components Normalization

In order to apply the principal components normalization, in which each speaker's vowels are expressed relative to his own principal axes, the polarity of corresponding axes of different speakers must agree. However, large differences in the directions of corresponding eigenvectors of the ten speakers make a consistent assignment of polarity beyond I_3 impossible, thereby limiting the use of individual principal components to the first three components, $(I_1, \ldots, I_3)$. Components of the cepstrum, $(C_4, \ldots, C_8)$, are used in place of the higher order principal components. We verified by inspection that C_4 is not highly correlated with I_3.

In speaker-independent recognition, the unnormalized mel-based cepstrum excluding C_0 achieves a recognition rate of 68 percent correct ($F = 2.1$), compared with a rate of 58 percent correct ($F = 1.7$) for the principal components normalization using $(I_1, \ldots, I_3, C_4, \ldots, C_8)$. To check whether the observed effect is due to removal of the overall energy component from the cepstrum, we equalized the overall energy in all log channel energy spectra prior to computation of the covariance matrix and subsequent processing. This gave even worse results: 49 percent correct recognition, $F = 1.0$.

4.2 Normalization by Vowel Space Centering

Since the principal components approach by itself appears to lack promise, we turned to another technique. One way in which vowel spaces tend to differ is in overall position relative to the origin of the parameter space. Centering normalizations attempt to remove this aspect of between-speaker variability by translating each speaker's vowel system appropriately. Arbitrarily, we choose the origin of the parameter space as the common center and translate all speakers' vowel systems so that their means occur at the origin. The potential benefit of such a centering operation has been shown in the log spectral domain.[8] The operation we propose is closely related to these, but is expressed in the transform domain.

[6] For details, see S.B. Davis and P. Mermelstein (1980), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. on Acoust. Speech and Signal Processing* **ASSP-28**(4), pp. 357–366.

[7] The global covariance matrix is computed over all spectral frames of all tokens from all speakers. The principal components transformation represents each token in the coordinate system defined by the eigenvectors of the global covariance matrix, in order of decreasing eigenvalue.

[8] L.C.W. Pols (1977), *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words*. Soesterberg (The Netherlands): Institute for Perception TNO, pp. 82-83. Also, M.J. Hunt (1981), "Speaker adaptation for word-based speech recognition systems," *Journ. Acoust. Soc. Am.* 69(S1), pp. S41–S42.

The centering operation consists of calculating, for each speaker, the average parameter vector over all tokens of all phonemes, weighting each token equally.[9] Normalization is then achieved by subtracting each speaker's mean parameter vector from each of his token vectors, thus "centering" his vowel space about the origin. The first two data columns of Table 1 compare recognition accuracies of unnormalized versus centered parameter vectors for three of the parameter sets previously discussed. Columns four and five show similar results for F. For all three parameter sets, cepstrum, global principal components, and individual principal components, centering causes a dramatic improvement in both recognition and F-ratio. The best parameter set still appears to be the cepstrum excluding C_0.

4.3 Normalization by Scaling of Dimensions

Vowel space centering causes dramatic improvements in both recognition performance and F-ratio. Can additional transformations further improve performance? Since centering has eliminated variability in the *position* of the vowel space center, one might speculate that eliminating interspeaker variability in the overall *range* of the parameters may also help.

For each of the ten speakers, the variance of each parameter is calculated over all tokens of all phonemes. Each parameter of each of the nine training speakers is then scaled so that its variance matches that of the test speaker's corresponding parameter. This technique preserves the natural weighting of the various parameters while equalizing variance in each dimension across all speakers. After centering and scaling, training data were averaged to form fifteen phoneme templates. Table 1 shows the results of scaling for the three parameter sets under discussion. Scaling increases the performance of $(C_1, \ldots, C_8)$ from 79 percent to 81 percent correct.

5.0 Conclusions

The 88 percent speaker-dependent recognition rate achieved using parameters $(C_1, \ldots, C_8)$ equals the rate with which human listeners perform an equivalent vowel classification task as reported by Eskenazi and Liénard.

In all the experiments we have performed, normalization effectiveness of principal components as measured by F-ratio and recognition accuracy never exceeds that of the mel-based cepstrum excluding C_0. This may be because of the rather small amount of vowel data used to determine each speaker's covariance matrix (about 13 to 14 seconds per speaker). In order to be useful as a speaker-adaptive technique, however, a speaker normalization method should show improvement with small amounts of data. We can conclude that the straightforward application of principal components as a normalization technique is ineffective.

On the positive side, we conclude that for all parameter sets investigated, centering is a powerful and effective method for normalization of vowels. Once centering has been applied, a small additional benefit may be derived from scaling. The reason centering is so effective appears to be its ability to normalize interspeaker differences in glottal spectrum shape (voice quality).

The best speaker-independent recognition score achieved for any 8-dimensional parameter set was 81 percent. When D is increased to 10, the same speaker-independent technique achieves 83 percent correct, as shown in Table 2, but does not improve significantly as D is further increased. The 83 percent speaker-independent result begins to approach the 88 percent human performance reported by Eskenazi and Liénard.

[9] Similar results are obtained by weighting each phoneme class equally.

	Percent Correct			**F-ratio**		
Parameter Set	**Unnormalized**	**Centered**		**Unnormalized**	**Centered**	
		Unscaled	**Scaled**		**Unscaled**	**Scaled**
$(C_1, \ldots, C_8)$	68 %	79 %	81 %	2.1	5.0	5.3
$(G_1, \ldots, G_8)$	57 %	75 %	78 %	1.4	4.6	5.7
$(I_1, \ldots, I_3, C_4, \ldots, C_8)$	58 %	72 %	74 %	1.7	4.5	4.9

Table 1. Recognition rates and F-ratios for three parameter sets (mel-based cepstrum, global principal components, and individual principal components) under three different normalization conditions: unnormalized, centered but not scaled, centered and scaled.

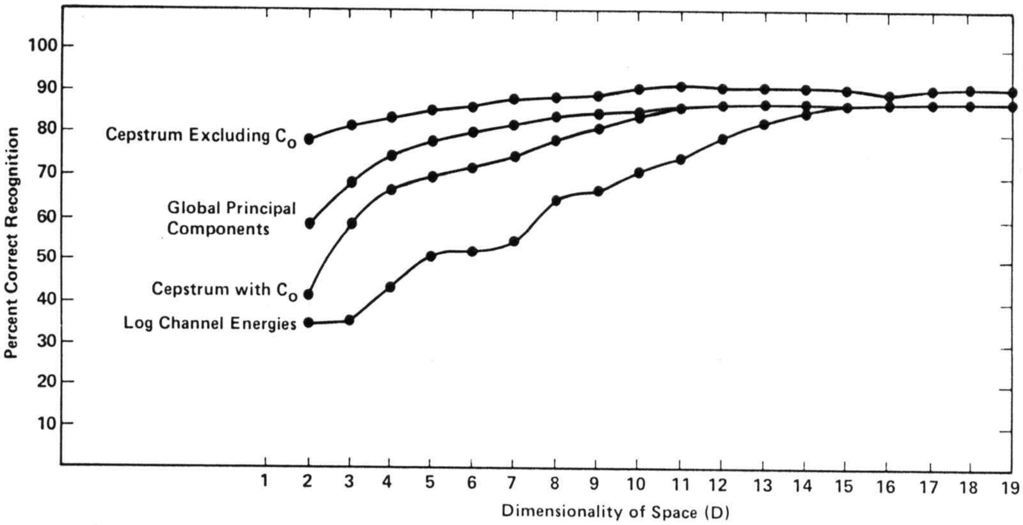

Fig. 1. Recognition scores in speaker-dependent experiments for four parameter sets: cepstrum excluding C_0, cepstrum with C_0, global principal components, and log channel energies. Number of parameters, D, varies from 2 to 19.

	/i/	/e/	/ε/	/y/	/œ/	/ø/	/a/	/*a*/	/⊃/	/o/	/u/	/ε̃/	/œ̃/	/ã/	/õ/	Percent Correct
/i/	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100 %
/e/	2	47	5	1	0	0	0	0	0	0	0	0	0	0	1	84 %
/ε/	0	8	55	0	6	1	1	1	0	0	0	0	2	0	0	74 %
/y/	1	0	0	17	0	0	0	0	0	0	0	0	0	0	0	94 %
/œ/	0	0	2	0	25	2	0	0	1	0	0	0	1	0	0	81 %
/ø/	0	3	2	0	2	32	0	0	0	0	0	0	0	0	1	80 %
/a/	0	0	1	0	0	0	78	37	0	0	0	2	2	0	0	96 %
/*a*/	0	0	0	0	0	0	12	49	0	0	0	2	0	5	0	90 %
/⊃/	0	0	0	0	13	0	0	3	69	3	0	6	1	10	2	64 %
/o/	0	0	0	0	0	0	0	0	0	91	1	0	0	2	8	89 %
/u/	2	0	0	0	0	2	0	0	0	7	117	0	0	1	6	87 %
/ε̃/	0	0	1	0	2	0	2	0	4	0	0	7	4	1	0	52 %
/œ̃/	0	0	0	0	4	0	3	1	1	0	0	11	6	3	0	59 %
/ã/	0	0	0	0	0	0	0	0	0	0	0	0	1	27	2	90 %
/õ/	1	0	0	0	0	1	0	0	0	3	1	0	0	2	31	79 %

Overall speaker-independent recognition rate: 83 %

Table 2. Confusion matrix for 83 % speaker-independent recognition rate (excluding confusions between /a/ = /*a*/ and /œ̃/ = /ε̃/, shown inside boxes) achieved using mel-based cepstrum parameters ($C_1, \ldots, C_{10}$) centered by speaker's token mean and scaled by speaker's token variance. $N = 891$. F-ratio $= 4.9$.