DIMENSION REDUCTION OF ACOUSTIC VOWEL DATA

Mylène Pijpers, Michael D. Alder and Roberto Togneri
Centre for Intelligent Information Processing Systems
Department of Electrical and Electronic Engineering
Department of Mathematics
The University of Western Australia

Abstract

From 33 speakers in the Timit database a total of around 88 vowel utterances was extracted. These represent eight vowel categories. Each waveform segment was processed, first by taking a FFT on 32 msec frames, then by binning into 12 mel-spaced frequency bands. This way each frame is described by 12 numbers, hence it becomes a point in $\mathbb{R}^{12}$. Each vowel utterance is a series of points and a short trajectory in $\mathbb{R}^{12}$. The eight vowel classes become eight clusters of points in the speech space.

The covariance matrix and the centers of the clusters were computed and dimension estimates of each vowel cluster by Principal Components Analysis show that the centers of the eight vowel clusters are all situated close to a plane and their principal axes make a small and consistent angle with respect to this plane. This confirms the results of Plomp e.a. (1969), who found the vowel space to be essentially 2 dimensional. Projecting the centers of the vowel clusters to the plane gives a representation of the vowels which is very similar to the conventional F1-F2 plot and the vowel diagram.

1 Introduction

1.1 Description of vowel sounds

Vowel sounds have been described in terms of perceptual differences (Pols e.a., 1969) and differences in their articulatory (Plomp e.a., 1969) and physical properties (Pols e.a., 1973). The description of the physical properties of speech sounds is done by describing the spectrum of the acoustic speech signal at a certain point in time. This can be done by determining the formants in the spectrum (Pols e.a, 1973) by finding LPC coefficients (Makhoul, 1975) or by computing the filterbank coefficients (Plomp e.a., 1969).
Filterbank coefficients to describe the physical vowel properties give results that can very easily be used for automatic speech recognition. Description of the speech space by LPC coefficients can be considered to be a continuous transformation of the 12 dimensional filterbank space we use in this paper (Togneri e.a., 1992).

1.2 The vowel data

We used vowel segments from 33 males in the TIMIT speech database CD-rom (DARPA, 1988). Eight vowels were chosen for further analysis, and each utterance was checked for clicks and random noises. Each of the resulting vowel groups contains 80 to 140 utterances, with a duration between 0.05 and 0.2 seconds In the notation used for the TIMIT database, followed by IPA notation, the vowels are the following: AA /ɑ/, AE /æ/, AH /ʌ/, AO /ɔ/, EH /ɛ/, IH /I/, IY /iʸ/ and UX /y/.
Vectorization of the speech fragments to 12 dimensional simulated filterbank coefficients was done, in C on SUN Sparc2 workstations. A FFT over 512 points with a 200 point advance was applied to the speech data, previously digitized at a frequency of 16 KHz. For every 512 point frame with a duration of 32 milliseconds the resulting values were reduced to 12 filterbank coefficients using a simulated filterbank of 12 overlapping filters. The filters are distributed on a Mel scale over the frequency range from 0 to 5 kHz.
Each vowel utterance, as it has been described above, is now represented in 12 filterbank values per frame of 32 msec. Each frame can be looked upon as a point in a 12 dimensional representation space. A series of frames, belonging to the same utterance, can be seen as describing a track through this representation space. The frames of several utterances of the same vowel spoken by different persons, form overlapping clusters in the 12 dimensional space. The centroid of each cluster, the means of all the frames for one vowel, is regarded as representing the location of the vowel in the 12 D space.

2 Method: Structure of the vowel data

2.1 Calculations on the data

To describe a cluster of points in the representation space the covariance matrix of all the points may be computed. Next the eigenvalues and eigenvectors of the cluster can be computed.
The number of significantly large eigenvalues of the cluster denotes the number of important dimensions, and their relative importance, within the representation space. These are called the principal components of the cluster. The square root of the eigenvalue gives the length of the radius of the hyperellipsoid along the corresponding axis of the subspace. These corresponding axes are the eigenvectors and they indicate the positioning of the cluster in the representation space, the shape of the hyperellipsoid is described by the eigenvalues.

2.2 Fitting a plane using principal components analysis

By taking two orthogonal vectors in the 12 dimensional simulated filterbank space we define a plane in that space. Projecting all the datapoints onto this plane will give a two-dimensional view through the space [3]. Let $\mathbf{u}$, $\mathbf{v}$ be two orthogonal vectors in $\mathbb{R}^n$, i.e. $\mathbf{u}\cdot\mathbf{u} = \mathbf{v}\cdot\mathbf{v} = 1$ and $\mathbf{u}\cdot\mathbf{v} = 0$.
Then for every $\mathbf{x} \in \mathbb{R}^n$ $(\mathbf{x}\cdot\mathbf{u}, \mathbf{x}\cdot\mathbf{v})$ is the projection of $\mathbf{x}$ onto the plane spanned by $\{\mathbf{u}, \mathbf{v}\}$. By operation on the window $\{\mathbf{u}, \mathbf{v}\}$ by an orthogonal n×n matrix we

may view the points in $\mathbb{R}^n$ from different orientations. By changing the position of the plane, according to computed best-fit vectors, or by rotating it at random in the 12 dimensional space, we obtain different views of the clusters in the 12 dimensional space.
To find the projection plane that gives the best discrimination possible between the centroids of the clusters, we compute the Minimized Squared Error (MSE) plane for the centroids. With the relatively low number of 8 different vowels we get 8 centroids in the 12 dimensional representation space. It is obvious these points must occupy a subspace from at most 7 dimensions. Possibly this subspace has even much lower dimensions. To find out if it is possible to fit a plane in the 12 dimensional representation space through the 8 centroids of the vowel clusters we must minimize the mean of the squared distances between all centroids and the proposed plane. This is done by computing *their* covariance matrix. This means we have to find the eigenvalues and eigenvectors of this cluster of 8 points. The proportion of the eigenvalues indicates the importance of the respective dimensions of the multidimensional object in the 12 dimensional representation space. The 2 eigenvectors that correspond with the 2 largest eigenvalues describe the MSE plane for the 8 points.

3 Results: Dimension of the vowel data

3.1 Relative positioning of the vowel clusters

In Figure 1. a graph shows the square root of the eigenvalues in descending order for each vowel cluster. The square root of the eigenvalue represents the length of the radius for that dimension of the hyperellipsoid. The 12 eigenvectors form an orthogonal basis for a 12 dimensional space that is located within the 12 dimensional representation space. For each of the 8 clusters this new basis is different, "adapted" to the best fit on the cluster.

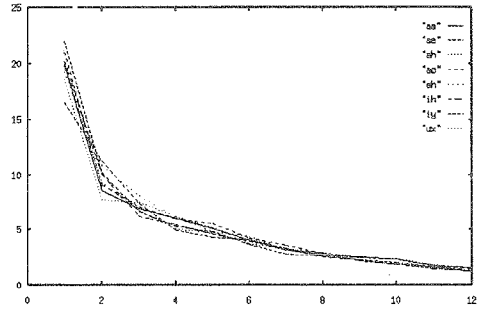

Figure 1. The decreasing values of the square root of the 12 eigenvalues for each vowel cluster.
However, if we compare the most important dimensions it can be seen that the vowel clusters are located in similar positions in the representation space. The direction of the first eigenvector of each cluster, the first principal component, is quite similar for all clusters, and the angles between them are between 5 and 19 degrees. This is a very small angle in a 12 dimensional space.

The shape of the different vowel clusters is also similar. The radius of the first dimension (the square root of the largest eigenvalue) is on average twice the radius of dimension 2. Dimension 2 and 3 are often very close and might easily be interchanged had the data been slightly different. The same applies to the eigenvalues of the subsequent dimensions, they decrease slowly as can be seen in Figure 1. which shows the decrease of the square root of the eigenvalues for each vowel cluster.

3.2 Eigenvalues and eigenvectors of the 8 centroids, the MSE plane

For the cluster that only consists of the 8 centroids of the 8 vowel clusters, in the 12 dimensional representation space, the covariance matrix and eigenvalues and eigenvectors can also be computed Table 1. lists the 8 centroids, in 12 dimensions. Showing their values in the 12 filterbank region spanning the frequency range from 0 to 5 kHz.

Filter	AA	AE	AH	AO	EH	IH	IY	UX
1	71.48	70.88	71.87	69.98	71.70	70.94	70.66	70.09
2	79.32	78.94	81.00	79.08	81.74	81.39	78.97	78.75
3	88.19	85.58	87.41	86.81	86.46	79.16	71.90	71.18
4	88.28	81.85	82.62	85.13	78.44	69.46	63.40	62.79
5	84.01	74.13	79.67	77.99	72.96	64.01	57.95	59.21
6	78.48	74.33	76.83	65.62	73.54	64.50	56.67	61.58
7	68.67	75.17	67.77	57.74	74.55	69.74	61.72	66.22
8	64.77	69.65	63.04	56.81	70.03	68.99	68.55	66.45
9	62.83	66.70	62.66	56.75	67.71	66.75	68.38	61.91
10	59.76	60.70	60.93	55.12	63.14	61.51	63.95	57.93
11	57.23	59.05	58.58	52.31	61.23	60.12	61.73	57.15
12	49.38	51.56	49.88	44.89	52.08	49.82	49.96	44.85

Table 1. The centroids of the vowel clusters, in the 12 dimensional filterbank spacespanning the frequencies from 0 to 5 kHz.

The 7 values for the squared root of the eigenvalues computed from the cluster of centroids are: 16.18 8.60 3.17 2.09 1.58 0.50 and 0.19. The first 2 eigenvalues are relatively large and the third and fourth dimension could have some influence, whereas the other 3 could easily be neglected.

3.3 Angles between principal axis of vowels and the MSE plane

The smallest angle between the MSE plane for the 8 means of the vowel clusters and the first eigenvector of each separate vowel cluster lies between 19 and 27 degrees, with a mean of 22 degrees. For every vowel cluster this smallest angle can be found in Table 2.

To find the largest angle θ between a vector $\mathbf{c}$ and a plane, the latter defined by 2 orthogonal vectors $\mathbf{a}$ and $\mathbf{b}$, the best fit angle α has to be found for which the vector $\cos\alpha$ vector $\mathbf{a}$ + $\sin\alpha$ vector $\mathbf{b}$ gives the largest angle θ with vector $\mathbf{c}$.

AA	AE	AH	AO	EH	IH	IY	UX
21	24	22	19	27	22	21	22

Table 2. The angle between the first axis of the hyperellipsoid of the vowel clusters and the plane formed by the first and second eigenvector of the means of the vowel clusters.

3.4 Projecting the centroids of the vowel clusters to a plane

In Figure 2. the centroids of the vowel clusters are projected onto the plane spanned by the first two eigenvectors of the cluster of centroids, which is the Mean Squared Error plane mentioned before.

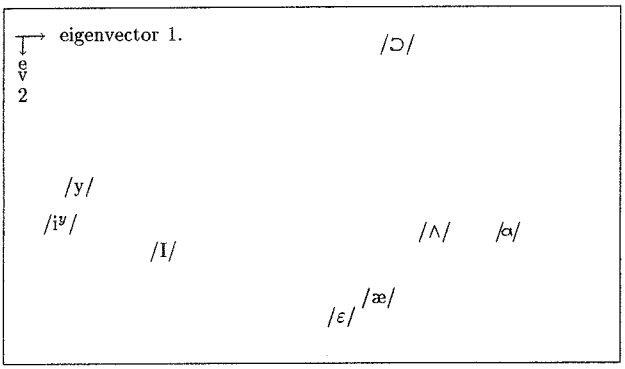

Figure 2. The diagram showing the projection of the 8 centroids onto the plane defined by the first two eigenvectors (the MSE plane).

4 Conclusions

4.1 The centroids of the vowel clusters in the speech space

Projecting the 12 dimensional datapoints to the 2 dimensional MSE best fit plane shows the configuration in Figure 2.
The vowel space can sufficiently be described as a flattish 3 dimensional object. A diagram that includes the third dimension suggests a possibly wedge shaped space but to confirm this more different vowels, covering the whole range, should be added to the speech data.
Projection to the plane determined by the first two eigenvectors of the cluster of centroids shows a strong similarity to the plot of the first two formants, F1 and F2, of the vowels. This is not really surprising, because F1 and F2 of the vowel indicate the peaks in the spectrum, which, in turn, will generate the largest eigenvalues. The projection to the plane given by the first and third eigenvector, however, does not look like the plot of F1 and F3.

4.2 The vowel clusters in relation to the MSE plane of the centroids

Considering the relatively small angles between the first eigenvectors of the vowel clusters we may conclude the vowel clusters are parallel positioned in the 12 dimensional representation space. The first principal axis of the clusters could indicate the loudness of the speech signal.
As can be seen in Table 2. the angles between the principal components of the clusters and the plane are very similar, we might visualise this as the MSE plane with 8 ellipsoids, the vowel clusters, lengthwise sticking out of it.

4.3 Automatic Speech and Speaker Recognition

Speculating as to what the several principal components could be representing we think that the first eigenvector could well determine the loudness of the speech signal. Some other eigenvectors can indicate the place of the cluster relative to the vowel plane. And maybe the distinction between different speakers can be made using only a few of the principal components. The techniques described can be used in automatic speech recognition to classify vowels. Similar techniques have been tested for vowel recognition. The same strategies as described above could be used for consonants. The properties of the vowel clusters, having more important eigenvalues and eigenvectors than the cluster of centroids, may be of use to design automatic speaker recognition systems.

5 References

DARPA TIMIT *Acoustic Phonetic Continuous Speech Database* , December 1988.
Makhoul J. *Linear Prediction: a Tutorial Review*, Proc. IEEE, vol.63, 1975 pp561-580.
Pijpers M. and Alder M.D. *Affine Transformations of the Speech Space*, Fourth Australian Conference on Speech Science and Technology, Brisbane, 1992, pp124-129.
Plomp R., Pols L.C.W. and van de Geer J.P. *Dimensional Analysis of Vowel Spectra*, J.Acoust.Soc.Am. Vol 41, 1967 pp707-712.
Pols L.C.W., van der Kamp L.J.Th. and Plomp R. *Perceptual and Physical Space of Vowel Sounds*, J.Acoust.Soc.Am. Vol 46, 1969 pp458-467.
Pols L.C.W., Tromp H.R.C. and Plomp R. *Frequency analysis of Dutch vowels from 50 male speakers*, J.Acoust.Soc.Am. Vol 53, 1973 pp1093-1101.
Togneri R., Alder M.D., and Attikiouzel, Y. *Dimension and Structure of the Speech Space*, IEE Proc-I, Vol 139, 1992 pp123-127.