# A NEW PROPOSAL FOR METRIC IN PERCEPTUAL MULTIDIMENSIONAL SCALING

Ondrej Šuch[1], Štefan Beňuš[2]

[1]University of Žilina and Mathematical Institute of Slovak Academy of Sciences, Slovakia

[2]Constantine the Philosopher University and Institute of Informatics of Slovak Academy of Sciences, Slovakia

ondrej.such@gmail.com and sbenus@ukf.sk

## ABSTRACT

We analyze a simplified probability model that assumes normality and homoscedasticity of vowel parameters' distributions. We arrive at explicit formulae that describe functional relationship between a natural Euclidean distance function in the model and the confusability of vowel categories. We propose a solution to address a common case when confusability of a pair of vowel categories is very low, which leads to uncertainty of distance value. The solution is applicable when the class boundary is provided by a real-valued discriminating function. Examples of perceptual vowel diagrams obtained with new metric are presented based on vowel samples spoken by male speakers from TIMIT corpus.

**Keywords:** multidimensional scaling, perceptual vowel space, TIMIT, annotated corpus

## 1. INTRODUCTION

Vowels constitute an important set of speech gestures in any language. Better understanding of a language, or a language corpus can be gained by visualizing the vowel set. This can be done in three approaches, articulatory, acoustic and perceptual. One of the earliest diagrams describing vowel space is the IPA chart, which maps vowels across two essentially articulatory dimensions, openness and backness. Another common way to describe the vowel space is via formants, especially through an inverted F1-F2 plot. Formants are measurable acoustic qualities of sound and thus have an added advantage of being exact. In our paper we are concerned with the third approach, namely visualization of perceptual vowel space.

Perception usually refers to human perception of the sound. Many studies have been conducted via psychoacoustic experiments where listeners are asked to compare vowel qualities. Often, multidimensional scaling (MDS) [1] was then used to describe the perceptual vowel space [2–9]. Generally, the resulting vowel configuration mirrored IPA chart and the dimensions could be given a phonetical meaning, e.g. backness, openness, rhotacization or F0.

With the advancement of machine learning that brought speech recognition to daily lives of many people, perception of speech is no longer an exclusive domain of humankind. Since MDS was used with great success in human studies, it is natural to ask how to extend tried-and-true MDS to visualize a machine's perceptual vowel space?

The key requirement to carry out MDS is to define a distance function between vowel categories. Such a function could be based on a distance function between individual pairs of vowels [10, 11]. We eschew such approach, since it is hard to make a strong argument for any particular vowel distance function. However, by imposing simplifying assumptions on the vowel space, in the following section we will be able to deduce the existence of a natural distance of vowel categories derived from vowel pair confusability (Theorem 1).

## 2. A THEORETICAL MODEL

The result of an MDS procedure is a low dimensional display of vowel categories. If we are interested in perceptual vowel chart, the key requirement is that vowels closer to each other in the chart imply "perceptual closeness". This "perceptual closeness" is directly related to confusability, or the likelihood of incorrect classification between pairs of vowels. Such likelihood is of course affected by prior probabilities $p(C_i)$ of vowel categories $C_i$. For simplicity througout the paper we assume a uniform prior $p(C_i) = p(C_j)$. We define the asymmetric confusability $\mathrm{aconf}(\cdot)$ of a binary recognizer $R$ for pairs of vowel categories $C_i, C_j$ as

$$\mathrm{aconf}(C_i, C_j) = p(R \text{ thinks } v \text{ is } C_i | v \text{ is } C_j) \quad (1)$$

and symmetric confusability $\mathrm{conf}(C_i,C_j) = \mathrm{conf}_R(C_i,C_j)$ by

$$\mathrm{conf}(C_i,C_j) = \frac{\mathrm{aconf}(C_i,C_j)+\mathrm{aconf}(C_j,C_i)}{2}. \quad (2)$$

Then a vowel category distance function that has the expected "closeness" property can be defined by setting

$$d(C_i,C_j) = \zeta(\mathrm{conf}(C_i,C_j)), \quad (3)$$

where $\zeta(s)$ is any decreasing function on $(0,\frac{1}{2})$. Now we will describe a simple probabilistic model of vowel space, where one choice for zeta function is quite natural.

Assume further that:

A1) any vowel can be specified using just two real valued parameters, call them $x$ and $y$,

A2) for a given vowel category $C_i$, latent vowel parameters $x$ and $y$ follow a normal distribution in two-parameter space with mean $\mathbf{m}_i$ and covariance matrix $\Sigma_i$.

A3) for two vowel categories, the Gaussians are homoscedastic, i.e. their covariance matrices are all the same, equal to $\Sigma$.

Under these assumptions, the optimal classifier between two vowel categories is just LDA and the confusability can be then explicitly determined. The LDA projection vector $\mathbf{p}_{ij}$ for classes $C_i$, $C_j$ is $\mathbf{p}_{ij} \propto \Sigma^{-1}(\mathbf{m}_i - \mathbf{m}_j)$. Let $\mathbf{m}_{ij}, \sigma_{ij}^2$ be the parameters of normal distributions of $C_i$ when projected onto $\mathbf{p}_{ij}$. Set $\Delta_{ij} = ||\mathbf{m}_{ij} - \mathbf{m}_{ji}||$.

The crucial point is that confusability between $C_i$ and $C_j$ depends only on the ratio

$$F_{ij} = \Delta_{ij}/(2\sigma_{ij}) \quad (4)$$

namely

$$\mathrm{conf}(C_i,C_j) = \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\int_0^{F_{ij}} e^{-t^2/2}dt \quad (5)$$

$$= \frac{1}{2} - \frac{1}{\sqrt{\pi}}\int_0^{F_{ij}/\sqrt{2}} e^{-w^2}dw \quad (6)$$

$$= \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{F_{ij}}{\sqrt{2}}\right)\right) \quad (7)$$

Going back to (3), we have the following result.

**Theorem 1.** *Assume that vowel categories satisfy A1)-A3) and $R(i,j)$ is an optimal classifier of vowel categories $C_i$ and $C_j$. If we set*

$$d(C_i,C_j) = c \cdot erf^{-1}\left(1 - 2conf_{R(i,j)}(C_i,C_j)\right), \quad (8)$$

*then there exist points $P_i$ in the plane with $d(P_i,P_j) = d(C_i,C_j)$.*

*Proof.* First note that invertible affine transforms of $x$-$y$ plane have no effect on the assumptions of our vowel space model. Moreover, $F_{ij}$ and confusability do not change either. It follows that we may assume without loss of generality that $\Sigma$ is the identity matrix. Then $F_{ij} \propto ||\mathbf{m}_i - \mathbf{m}_j||$ and we are done by (5)-(7). □

Corollary of the proof of the theorem is that by making this particular choice of $\zeta(s)$, MDS will recover the original position of $\mathbf{m}_i$ when $\Sigma = \mathbf{I}$. This is guaranteed, since MDS will find a two-dimensional configuration of points with given distances, if such configuration exists.

One practical problem when applying this theorem is that some pairs of vowels on the opposite sides of IPA chart can be discriminated quite well. This implies that empirical confusability is very near 0 or even 0, where the inverse error function approaches infinity quite rapidly. Our suggestion to address this problem is to avoid a detour through confusability function, whenever possible. Many popular classifiers, such as LDA or SVM produce a discriminating function $f$, whose evaluation is used for discrimination between classes. Instead of using formula (3), one may evaluate Fisher discriminant of $f$ which is defined as

$$D_{\mathrm{Fisher}}(C_i,C_j) = \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2}, \quad (9)$$

where $\mu_k$, $\sigma_k^2$ are means and variances of $f$ applied to samples from vowel category $C_k$. If $f$ is close to an optimal classifier between $C_i$ and $C_j$ then one has $D_{\mathrm{Fisher}}(C_i,C_j) \approx 2F_{ij}^2$ and thus one may set

$$d(C_i,C_j) = \sqrt{D_{\mathrm{Fisher}}(C_i,C_j)}. \quad (10)$$

Let us make a brief discussion on applicability of the assumptions A1-A3. The first one is an approximation supported by several studies e.g. [12], [13]. A second one is stronger, and likely implies that a vowel category contains only one distinguishable allophone. Finally A3) is the strongest, and likely to hold only under special circumstances. However MDS has been shown to produce phonetically meaningful results even in human trials under a lot of subjectivity. Thus even if A1-A3 are not perfectly satisfied, we may expect that using root of Fisher discriminant for vowel category distance function will yield valuable insights of the vowel space.

**Figure 1:** Result of Experiments 1 (left) and 2(right). MDS placements that best fit metric defined by (10).

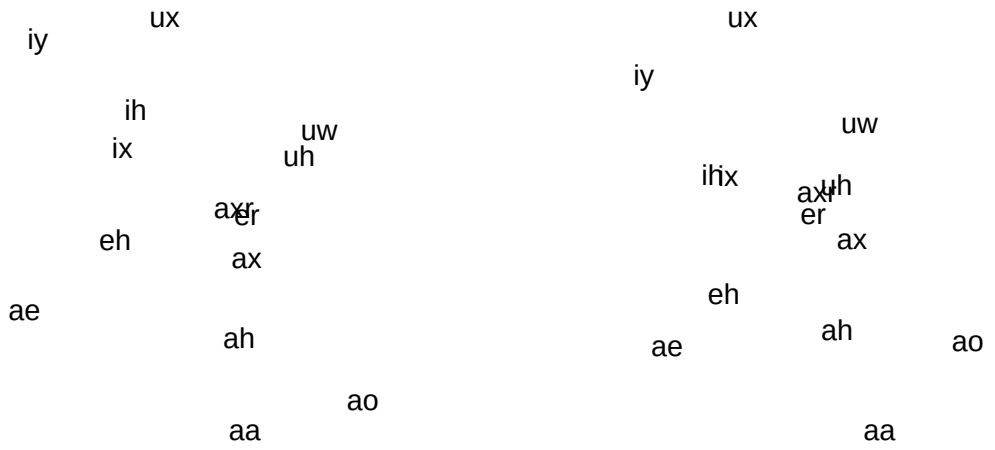
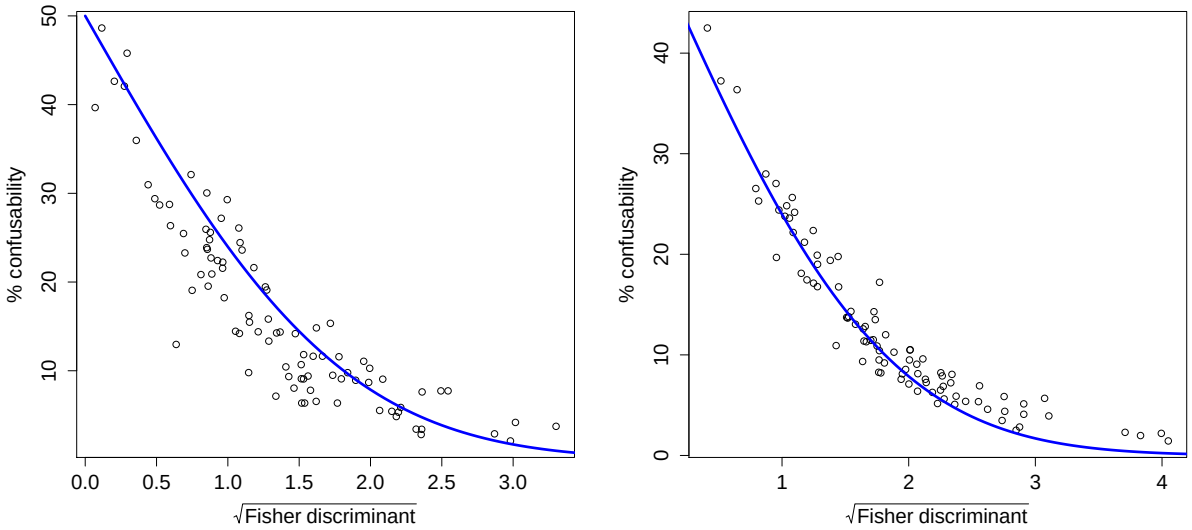



**Figure 2:** Empirical confusability of pairwise discriminating functions plotted against the square root of Fisher discriminant in Experiments 1 (left) and 2 (right). The curves show ideal correspondence predicted by (8)-(10).




## 3. EXPERIMENTS

We present the results of three experiments that outline applicability of the proposed MDS metric. All three experiments were carried out on vowels of TIMIT corpus of American English [14]. For our machine learning algorithm, whose "perceptual space" we wanted to visualize, we chose LDA. It is deterministic, fast to compute, resilient to over-training and optimizes Fisher discriminant, which we propose to use as MDS distance function (10).

### 3.1. Experimental details

LDA was trained on 2500 randomly chosen samples of male speakers from the corpus according to corpus annotation. In all cases the analysis window had width 512 samples and it was weighted with Han-

ning window.

**Experiment 1.** The grouping variable was TIMIT phoneme annotation. Groups correspondings to diphtongs and ax-h were excluded. The features provided to LDA were F1 and F2 frequencies in Hertz.

**Experiment 2.** As in Experiment 1, the grouping variable was TIMIT phoneme annotation. Unlike experiment 1, the features provided to LDA consisted of log periodogram. More precisely, 256 power spectral values for frequencies $k \cdot f$, where $k = 1, \ldots, 256$ and $f = 16000/512$ Hz.

**Experiment 3.** We considered only samples labeled 'aa' or 'ao'. There were 16 groups, two for each one of 8 geographical regions of the speaker. The features provided to LDA including F1-F4 frequencies in Hertz, their bandwiths in Hertz, log periodogram as in Experiment 2, logarithm of total power, and angles $\phi$, $\theta$ as defined in [13].

**Methods.** To carry out the experiments we developed custom code for R software. The code depended on packages fftw for Fourier transform, e1071 for Hanning window, phonTools for formant identification, and MASS for LDA.

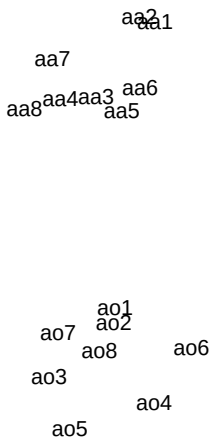
### 3.2. Discussion of experiments

Experiments 1 and 2 used the same training data, as well as the same testing data, thus allowing us to make a direct comparison of the results as presented in Figure 1. We can see that MDS found essentially the same configuration of points in the plane that represents the vowel categories. However, there are a few pairs of phonemes that have notably different distances, e.g. axr-er, uh-uw and ix-ih. Larger distances in Experiment 2 may be caused by extra power information available to the classifier, unlike in Experiment 1, where power information is not present. A common feature of the results is the emergence of distinctive aa-ao-uw-ux-iy-ae hexagon, contrasting with vowel trapezoid of IPA chart.

One may question applicability of Fisher discriminant in view of rather weak adherence of distribution of formants to homoscedasticity assumption A3). To that end we plotted confusability against the square root of Fisher discriminant in Figure 2 and found generally good correspondence of the two quantities with relationship predicted from (8) and (10).

In Experiment 3 one may expect that homoscedasticity assumption A3 is much more strongly satisfied than in experiments 1 and 2, since we consider the utterances of very similar vowels. Since finding subtle differences between regional accents is a rather delicate task we supplied the classifier with a variety of features as described above. The result found by MDS (Figure 3) to a large degree mirrors geographical relationships of speakers.

**Figure 3:** Result of experiment 3. MDS placement based on metric defined by (8). 1=New England, 2=Northern, 3=North Midland, 4=South Midland, 5=Southern, 6=New York City, 7=Western, 8=Army Brat)




## 4. CONCLUSION

The primary goal of the paper was to argue for a new metric to visualize perceptual vowel space. We showed a theoretical argument why and under what assumptions, the metric defined in (8) is the natural choice. The most notable feature of the experiments is the deformation of IPA vowel trapezoid to a "perceptual hexagon". It would be very interesting to test if neural correlates of the hexagon exist in the human brain [15–17].

Compared to F1-F2 plot, the principal advantage of the new method is that it is able to aggregate many more dimensions. The downside is the need to separately interpret phonetical meaning of dimensions and the inability to visualize any particular vowel instance in the MDS chart.

**Acknowledgment** O. Šuch was partially supported by grants University Science Park ITMS 26220220184 and APVV-0219-12. Š. Beňuš was supported by VEGA grant 2/0197/15.

## REFERENCES


[1] J. C. Gower. “Some distance properties of latent root and vector methods used in multivariate analysis”. In: *Biometrika* 53.3-4 (1966), pp. 325–338. DOI: 10.1093/biomet/53.3-4.325.

[2] R.A. Fox. “Individual variation in the perception of vowels: implications for a perception-production link”. In: *Phonetica* 39 (1982), pp. 1–22.

[3] R. A. Fox. “Perceptual Structure of Monophthongs and Diphthongs in English”. In: *Language and Speech* 26.1 (1983), pp. 21–60. DOI: 10.1177/002383098302600103.

[4] R. A. Fox. “Multidimensional scaling and perceptual features: evidence of stimulus processing or memory prototypes?” In: *J. Phonet.* 13 (1985), pp. 205–217.

[5] L. C. W. Pols, L. J. Th. van der Kamp, and R. Plomp. “Perceptual and Physical Space of Vowel Sounds”. In: *The Journal of the Acoustical Society of America* 46.2B (1969), pp. 458–467. DOI: http://dx.doi.org/10.1121/1.1911711.

[6] Brad Rakerd and Robert R. Verbrugge. “Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels”. In: *The Journal of the Acoustical Society of America* 77.1 (1985), pp. 296–301. DOI: http://dx.doi.org/10.1121/1.392393.

[7] R. Shepard. “Psychological representation of speech sounds”. In: E.E. David and P.B. Denes. *Human communication: a unified view*. 1972.

[8] Sadanand Singh and David R. Woods. “Perceptual Structure of 12 American English Vowels”. In: *The Journal of the Acoustical Society of America* 49.6B (1971), pp. 1861–1866. DOI: http://dx.doi.org/10.1121/1.1912592.

[9] R.A.Fox and M.D. Trudeau. “A multidimensional scaling study of esophageal vowels”. In: *Phonetics* 45 (1988), pp. 30–42.

[10] M. Huckvale. “ACCDIST: A metric for comparing speakers’ accents”. In: *Proceedings of the international congress of phonetic sciences*. Jeju, Korea, 2004, pp. 29–32.

[11] Emmanuel Ferragne and François Pellegrino. “Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics”. In: *Journal of Phonetics* 38.4 (2010), pp. 526–539. ISSN: 0095-4470. DOI: http://dx.doi.org/10.1016/j.wocn.2010.07.002.

[12] D. J. Broad and H. Wakita. “Piecewise-planar representation of vowel formant frequencies”. In: *The Journal of the Acoustical Society of America* 62.6 (1977), pp. 1467–1473. DOI: http://dx.doi.org/10.1121/1.381676.

[13] R.E. Turner and R.D. Patterson. “An analysis of the size information in classical formant data: Peterson and Barney (1952) revisited”. In: *J. Acoust. Soc. Jpn.* 33 (2003), 585â–589.

[14] John Garofolo et al. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. 1993. URL: https://catalog.ldc.upenn.edu/LDC93S1.

[15] C. Pantev et al. “Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields”. In: *Electroencephalography and Clinical Neurophysiology* 69.2 (1988), pp. 160–170. ISSN: 0013-4694. DOI: http://dx.doi.org/10.1016/0013-4694(88)90211-8.

[16] Jonas Obleser et al. “Cortical representation of vowels reflects acoustic dissimilarity determined by formant frequencies”. In: *Cognitive Brain Research* 15.3 (2003), pp. 207–213. ISSN: 0926-6410. DOI: http://dx.doi.org/10.1016/S0926-6410(02)00193-3.

[17] Anna Shestakova et al. “Orderly cortical representation of vowel categories presented by multiple exemplars”. In: *Cognitive Brain Research* 21.3 (2004), pp. 342–350. ISSN: 0926-6410. DOI: http://dx.doi.org/10.1016/j.cogbrainres.2004.06.011.