Intrinsic-cum-extrinsic normalization of formant data of vowels

Ananthapadmanabha T. V.Ramakrishnan A. G.

Citation: J. Acoust. Soc. Am. **140**, EL446 (2016); doi: 10.1121/1.4967311
View online: http://dx.doi.org/10.1121/1.4967311
View Table of Contents: http://asa.scitation.org/toc/jas/140/5
Published by the Acoustical Society of America

Intrinsic-cum-extrinsic normalization of formant data of vowels

T. V. Ananthapadmanabha
Voice and Speech Systems, Malleswaram, Bangalore 560003, India
tva.blr@gmail.com

A. G. Ramakrishnan
Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
ramkiag@ee.iisc.ernet.in

Abstract: Using a known speaker-intrinsic normalization procedure, formant data are scaled by the reciprocal of the geometric mean of the first three formant frequencies. This reduces the influence of the talker but results in a distorted vowel space. The proposed speaker-extrinsic procedure re-scales the normalized values by the mean formant values of vowels. When tested on the formant data of vowels published by Peterson and Barney, the combined approach leads to well separated clusters by reducing the spread due to talkers. The proposed procedure performs better than two top-ranked normalization procedures based on the accuracy of vowel classification as the objective measure.

[DDO]
Date Received: September 21, 2015 **Date Accepted:** July 14, 2016

1. Introduction

Formant frequencies measured over the mid-part of a vowel of American English, spoken in the same context (/hVd/) by talkers of different age and gender, unanimously labelled by native listeners, show a considerable spread in the F_2 versus F_1 space.[1] This has motivated researchers to look for a suitable transformation or normalization of the measured raw formant data to bring out the underlying invariance of vowels. The normalization is expected to reduce the spread in the formant data arising due to the influence of talker's gender and age, while preserving the relative mean positions of the vowels as in the original formant space.[2,3]

There is a huge amount of literature on vowel normalization, spanning over six decades, inhibiting a critical review in this short paper. We cite some secondary sources. Adank[4] gives a review of the literature up to 2003. The effectiveness of some select vowel normalization methods have been compared based on certain objective criteria.[5–7] Carpenter and Govindarajan[8] give a brief description as well as an evaluation of 32 intrinsic and 128 extrinsic procedures for the vowel classification task. Normalization in the context of sociolinguistics has also been reported.[7,9]

Some important milestones of research in this area are briefly covered. On an average, the vocal tract length (VTL) of an adult female (or child) is shorter than that of an adult male. Theoretically, this implies that all the formants be scaled inversely as the ratio of VTLs. However, the ratio of the mean formant frequency of adult female speakers to that of male speakers is both vowel and formant dependent,[10] varying over a wide range of 1.03 to 1.30 for the data published by Peterson and Barney[1] (abbreviated as P&B). This wide range combined with the fact that the mean formant frequency of adult female speakers has been reported to be lower than that of adult male speakers for a specific Swedish vowel,[10] has led researchers to speculate that factors other than VTL, such as possible gender based differences in articulation, may also contribute to the noted differences in the formant ratios.[11] F0 has also been considered to be an additional parameter for disambiguating vowels. For normalization, researchers have proposed differences such as $(F_1 - F0)$, $(F_2 - F_1)$, $(F_3 - F_2)$ and the ratios (F_2/F_1), (F_3/F_2), etc., in various frequency scales such as Koenig, log, mel, or Bark.[12,13]

The topmost performing normalization procedure for automatic vowel classification yields only about 80% accuracy even with the controlled context of P&B data.[8] Despite the availability of a large number of procedures, a fully satisfactory solution for normalization is yet to emerge.[6] This has motivated us to propose an intrinsic-cum-extrinsic normalization procedure, resulting in what we refer to as de-normalized formants. The effectiveness of the combined procedure in reducing the influence of

talker's age and gender is illustrated using the P&B data. Vowel classification using the pooled de-normalized formant values of all speakers (adult male, adult female, and child) is shown to give a very high accuracy (95%). The performance of the proposed procedure compares well with, or is better than, two top-ranked normalization procedures.[4,5]

2. Proposed method

2.1 Intrinsic normalization

The geometric mean of the first three formant frequencies[14,15] of a speaker's vowel sample is given by

$$GM123 = [F(1)F(2)F(3)]^{(1/3)}, \tag{1}$$

where $F(i)$ corresponds to the ith raw formant frequency in Hz. Let $AM(i)$ and $AF(i)$, $i = 1, 2, 3$ denote the mean values of the first three formant frequencies of adult males and females, respectively. Assuming $AF(i) = \alpha AM(i)$, the ratio of geometric means, GM123(female)/GM123(male) is equal to α. Hence GM123 may be expected to normalize any uniform scaling of the formant frequencies arising due to gender and age. The normalized formant frequency[14,15] of a given vowel sample is given by the ratio

$$NF(i) = F(i)/GM123, \tag{2}$$

where the ratio $NF(i)$ is a dimensionless quantity. Equation (2) makes use of speaker-specific data of the first three formants of only the given vowel sample. Hence the procedure has to be strictly called "speaker-intrinsic, formant-extrinsic, and vowel-intrinsic" normalization.[5] Instead, for the sake of brevity, we refer to the procedure as intrinsic normalization.

GM123 has a wide range of about 644 Hz (vowel /u/ of an adult male speaker) to 1400 Hz (vowel /æ/ of the same speaker) for the P&B data, i.e., a factor of more than 2. However, for a given speaker, VTL varies only by about 10% for different vowels. The over-correction in intra-speaker normalization results in a distortion of the vowel space. Due to the very low value of GM123 for back rounded vowels, in the NF_2 versus NF_1 space, these vowels lie above vowel /ɑ/ along the /ɑ/-/i/ direction instead of lying below /ɑ/ in the /ɑ/-/u/ direction as in the raw formant space. In order to restore the original relative vowel positions, we propose an extrinsic de-normalization procedure.

2.2 Proposed extrinsic de-normalization procedure

Assumptions: In a normalization procedure, it is incorrect to assume the vowel identity of a sample to be known. Hence, the statistics of the formant data across all vowels, instead of vowel specific statistics, are used in the existing extrinsic procedures.[4–6,11] However, we make use of vowel specific statistics, the mean $\mu(i, j)$, and the standard deviation $\sigma(i, j)$ of vowel j. During the process of the proposed extrinsic normalization, the identity of the vowel sample is also determined. Since $\mu(i,j)$ and $\sigma(i, j)$ depend solely on a specific formant i of a specific vowel j, the procedure is "formant-intrinsic" and "vowel-intrinsic."[5] Since the statistics represent the average across speakers, it is "speaker-extrinsic." For the sake of brevity, we use the term "extrinsic."

Development of the proposed procedure: We define the geometric mean of the average formant frequencies for a given vowel as

$$GMA123 = [\mu(1)\mu(2)\mu(3)]^{(1/3)}. \tag{3}$$

Initially, we explored using the ratio GMA123/GM123 as the normalization factor in Eq. (2) instead of the reciprocal of GM123. The rationale is that while GM123 is expected to normalize for the inter-speaker differences, the factor GMA123 would restore the relative vowel positions. Further, the normalized values will now have the unit of Hz, with the range of values comparable to those of the raw formant data. However, both GM123 and GMA123 are common scale factors for all the three formants of a given vowel j. However, as noted in Sec. 1, formant ratios are both formant and vowel dependent. Hence we propose $\mu(i, j)$ itself as a scaling factor since it is both formant (i) and vowel (j) dependent.

Proposed extrinsic de-normalization: The intrinsically normalized formant values $NF(i)$ of a vowel sample are transformed to what we refer to as the de-normalized values. Since the vowel identity of a test sample is unknown, we use a "hypothesize-test" paradigm. Let V be the number of vowels in the database. We hypothesize the

index J, one at a time, of the unknown vowel and for each hypothesis J, the de-normalized formant value is determined as

$$DF(i,J) = NF(i) * \mu(i,J). \tag{4}$$

In our study, we find that the mapping from the dimensionless NF to DF with the unit in Hz does not affect the results.[16] Each vowel sample $NF(i)$ maps to V de-normalized values, $DF(i, J)$, for hypotheses $J = 1$, V of which only one hypothesis has to be selected. We test each hypothesis by computing the distance between the de-normalized first two formants and the mean values of the corresponding de-normalized formant data of the hypothesized vowel as

$$D(J) = Distance\langle DF(i,J), \bar{\mu}(i,J), \bar{\sigma}(i,J)\rangle, \quad i = 1,2, \tag{5}$$

where *Distance* $\langle\rangle$ denotes an appropriate distance measure (see Sec. 3.2). The third formant frequency has an indirect influence via $NF(i)$. Let $\bar{J}$ be the index for which $D(J)$ is the minimum. The vowel index is postulated as $\bar{J}$. Only $DF(i,\bar{J})$ is taken as the de-normalized value. That is, $NF(i)$ maps to $DF(i,\bar{J})$ in the de-normalized space. This procedure at once achieves vowel de-normalization as well as vowel classification.

A parallel to perceptual studies: Utilizing the mean and standard deviation values implies having *a priori* knowledge of the vowel space of a given language. The performance is known to degrade if anomalous information is given about the speaker's gender (male/female)[17] or the language (American English/Canadian English).[18] This suggests that a listener's performance of perceptual identification of vowels improves with *a priori* knowledge (or familiarity) of the talker's identity or gender or language. It is speculated that listeners use a "cognitive frame of reference" of the talker.[11] With this background, the use of *a priori* knowledge of the mean and standard deviation values of vowel formant data appears justified.

2.3 Experimental results and discussion

We have used the P&B data[19,20] for illustrating the procedure. There are 66, 56, and 30 samples for "men," "women," and "children" categories, respectively. We have considered all the (nine) vowels excluding the retroflex vowel /ɝ/. In the illustrations to follow, a vowel triangle[5,6,22] based on the mean values of the three corner vowels is also shown for the adult male and female speakers. Its relevance is discussed in Sec. 3.1. We have followed the convention used by P&B in selecting the orientation of the plot with vowel /u/ near the bottom-left of the graph. In all the figures, the same notation as given in Fig. 1 is followed.

A plot of raw formant data, F_2 versus F_1, is shown in Fig. 1. For the front vowels, the data show a wide spread across gender and age. Also, a considerable spread is seen within each vowel. The front vowels are not well separated and some back vowels (/ʊ/ and /u/, /ɑ/ and /ɔ/) heavily overlap. Also see Fig. 8 of Peterson and Barney[1] and Fig. 3 of Miller.[13]

In the de-normalized formant space, both the inter and intra speaker spread is reduced considerably (DF_2 versus DF_1 plot of Fig. 2). The relative positions of vowels

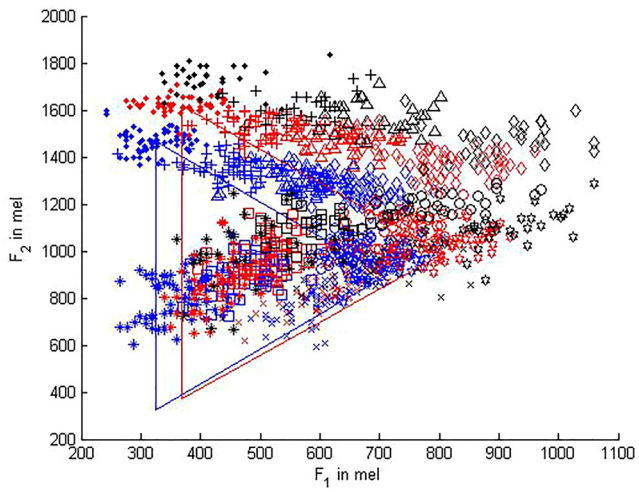

Fig. 1. (Color online) The plot of raw formant (F_2 versus F_1) data (Ref. 1) in mel scale. Filled dot: i, plus: ɪ, triangle: ɛ, diamond: æ, circle: ʌ, hexagon: ɑ, cross: ɔ, square: ʊ, and star, u. Blue: adult male, red: adult female, black: children.

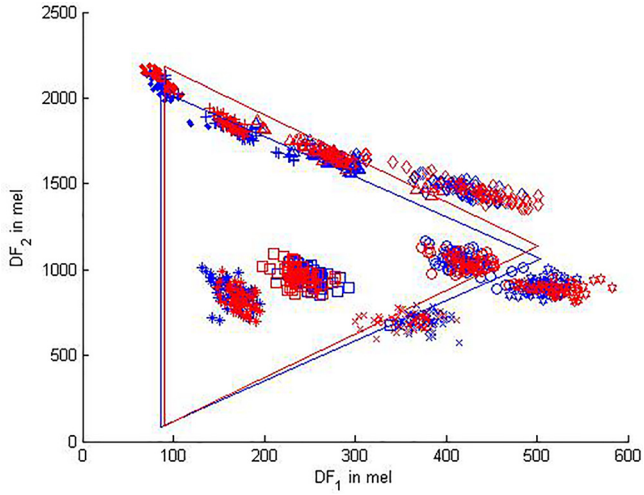

Fig. 2. (Color online) The plot of de-normalized formant (DF_2 versus DF_1) data in mel, obtained using the proposed procedure.

are preserved as in the raw formant data space. Tense/lax and high/low front vowels form distinct clusters. The separation amongst back vowels is surprisingly good. Clusters for vowels (/ʊ/ and /u/) and (/ɑ/ and /ɔ/) are also reasonably well separated.

3. Comparison with other methods

For comparison, we have chosen two top performing[6,7] normalization procedures, namely, the z-score[21] and the S-centroid.[5,6,22]

3.1 Formant plots and vowel triangles

The plots of formant values normalized using the z-score (Z_2 versus Z_1) and S-centroid (S_2 versus S_1) procedures are shown in Figs. 3 and 4, respectively. The spread in the data points arising due to gender and age difference is reduced for both the procedures. However, the vowel samples are widely scattered. In the case of S-centroid procedure, clustering is very good only for vowel /i/ as it acts as a reference corner. It is difficult to infer the number of vowels from the plots shown for z-score and S-centroid. In the de-normalized formant space, one distinct cluster per vowel is seen (Fig. 2).

One of the ways to study the effectiveness of a normalization procedure is to compare the overlap of vowel triangles for male (VTM) and female (VTF) speakers.[5,6,22] We give only a qualitative comparison. For the raw data (Fig. 1), VTF is much bigger than VTM and is significantly displaced upwards and to the right. For the proposed procedure (Fig. 2), VTF and VTM almost overlap except for a slight mismatch in the /i/-/ɑ/ direction. For the z-score normalization (Fig. 3), VTF is smaller than VTM with a slight mismatch in the /i/-/u/ direction. For the S-centroid method (Fig. 4), it is difficult

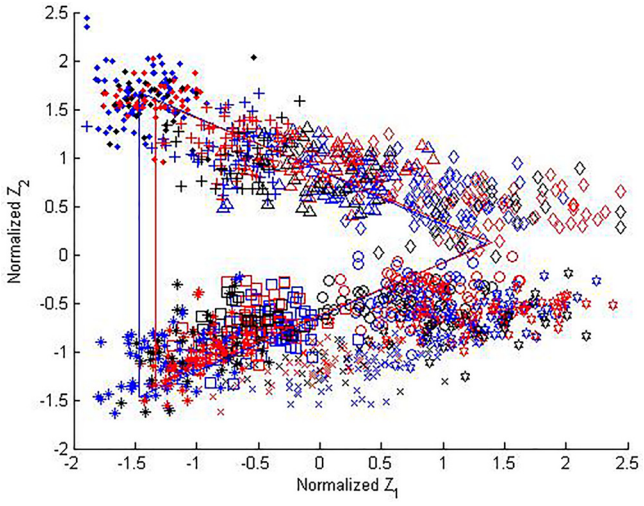

Fig. 3. (Color online) The plot of normalized formant data obtained using the z-score procedure (Ref. 21).

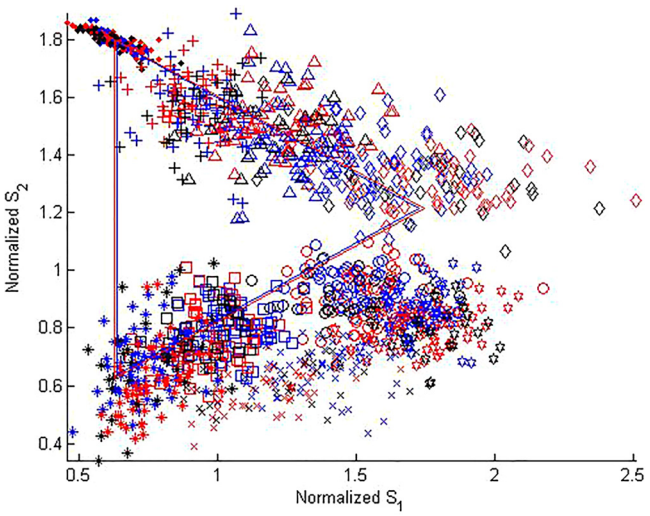

Fig. 4. (Color online) The plot of normalized formant data obtained using the S-centroid procedure (Ref. 5).

to discern the two vowel triangles as the overlap is almost complete. A vowel triangle is determined by only three normalized parameters, F_1 and F_2 of /i/ and F_1 of /ɑ/ and hence it does not reflect the spread of data. We propose to use the accuracy of vowel classification as an objective measure for a comparison of different normalization procedures.

3.2 Vowel classification accuracy as an objective measure

We assume a labeled database of formants of a given language to be available. The set of formant frequencies (F_1, F_2) in mel is used as the feature vector. The mean values $\bar{\mu}_1$ and $\bar{\mu}_2$ represent the vowel space. Given the test formant data, its nearest vowel in the vowel space is declared as the identity of the test vowel and compared with the known label. The overall accuracy for all the samples is determined. A similar procedure is applied on the normalized formant values of z-score and S-centroid procedures. Vowel classification is a part of the proposed procedure, as already noted in Sec. 2.2. We have used a weighted Euclidean distance (WED) measure given by

$$WED^2 = [F_1 - \bar{\mu}_1]^2/\bar{\sigma}_1^2 + [F_2 - \bar{\mu}_2]^2/\bar{\sigma}_2^2. \tag{6}$$

Selection of test samples: For the P&B database, for the gender-independent (MW) case, formant data of "men" and "women" categories and for the gender-age-independent (MWC) case, formant data of all the three categories are pooled together and used. Improvement in vowel classification accuracy, computed with the pooled normalized formant values over the accuracy obtained with the pooled raw formant data is considered as a measure of the effectiveness of the normalization procedure. The vowel dependent statistics ($\bar{\mu}$, $\bar{\sigma}$) are computed on the raw and normalized (or de-normalized) pooled formant data using the known labels. For automatic vowel classification, the statistics are to be computed from a training set.

Results: The classification accuracies for the raw data, S-centroid, z-score, and the proposed procedures are [82.9%, 85.0%, 85.7%, 95.2%] for the MW case and [77.2%, 84.5%, 84.4%, 94.9%] for the MWC case, respectively. The proposed procedure gives the highest accuracy of about 95%, nearly 10% higher than the S-centroid and z-score normalization procedures and 12% (18%) higher than the MW (MWC) case of raw data.

4. Conclusion

We have used vowel dependent statistics and proposed an intrinsic-cum-extrinsic procedure along with a "hypothesize-and-test" paradigm. For the given P&B database, the large spread observed in the acoustic space for different vowels and talkers has been effectively reduced. Clear clusters have emerged in the de-normalized formant space. The proposed procedure performs better than two top performing procedures in removing the influence of gender and age based on the accuracy of vowel classification as the objective measure. For future work, comparison with other procedures of normalization with rigorous objective measures may be undertaken and the applicability of the proposed procedure, over a larger database and in areas like sociolinguistics,

language change, influence of accent, etc., may be explored. The proposed procedure can also be applied on normalized data obtained with other procedures.

References and links

[1]G. Peterson and H. Barney, "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**(2), 175–184 (1952).

[2]S. F. Disner, "Evaluation of vowel normalization procedures," J. Acoust. Soc. Am. **67**(1), 253–261 (1980).

[3]D. J. L. Watt, A. H. Fabricius, and T. Kendall, "More on vowels: Plotting and normalization," in *Sociophonetics: A Student's Guide*, edited by M. Di Paolo and M. Yaeger-Dror (Routledge, London, 2010), pp. 107–118.

[4]P. Adank, "Vowel normalization: A perceptual study of Dutch vowels," Ph.D. thesis, University of Nijmegen (2003).

[5]A. H. Fabricius, D. J. L. Watt, and D. E. Johnson, "A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics," Lang. Var. Change **21**(3), 413–435 (2009).

[6]N. Flynn and P. Foulkes, "Comparing vowel formant normalization methods," in *Proceedings of ICPhS*, Hong Kong (2011).

[7]P. Adank, R. Smits, and R. V. Hout, "A comparison of vowel normalization procedures for language variation research," J. Acoust. Soc. Am. **116**(5), 3099–3107 (2004).

[8]G. A. Carpenter and K. K. Govindarajan, "Neural network and nearest neighbor comparison of speaker normalization methods for vowel recognition," in *ICANN'93, Proceedings International Conference on Artificial Neural Networks*, Amsterdam (1993).

[9]A. Fabricius, "Variation and change in the TRAP and STRUT vowels of RP: A real time comparison of five acoustic data sets," J. Int. Phon. Assoc. **37**(3), 293–320 (2007).

[10]G. Fant, *Speech, Sounds and Features* (MIT Press, Cambridge, 1973), Chap. 4.

[11]K. Johnson, "Speaker normalization in Speech perception," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. Remez (Blackwell Publishers, Oxford, 2005), pp. 363–389.

[12]A. Syrdal and H. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**(4), 1086–1110 (1986).

[13]J. D. Miller, "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. **85**(5), 2114–2134 (1989).

[14]J. D. Miller, A. M. Engebretson, and N. R. Vemula, "Vowel normalization: Differences between vowels spoken by children, women, and men," paper presented at the *100th meeting of the Acoustical Society of America*, Los Angeles (1980).

[15]H. M. Sussman, "A neuronal model of vowel normalization and representation," Brain Lang. **28**, 12–23 (1986).

[16]T. Kendall and E. R. Thomas, "The vowel manipulation and plotting suite," http://lingtools.uoregon.edu/norm/norm1.php (Last viewed November 4, 2016).

[17]K. Johnson, E. A. Strand, and M. D'Imperio, "Auditory-visual integration of talker gender in vowel perception," J. Phon. **27**, 359–384 (1999).

[18]N. A. Niedzielski, "The Effect of Social Information on the Phonetic Perception of Sociolinguistic Variables," Ph.D. dissertation, University of California–Santa Barbara, Santa Barbara, CA (1997).

[19]"Peterson Barney: Vowel formant frequency database," http://www.cs.cmu.edu/Groups/AI/areas/speech/database/pb/0.html (Last viewed November 4, 2016).

[20]R. L. Waltrus, "Current status of Perterson Barney vowel formant data," J. Acoust. Soc. Am. **89**, 2459 (1991).

[21]B. M. Lobanov, "Classification of Russian vowels spoken by different speakers," J. Acoust. Soc. Am. **49**(2), 606–608 (1971).

[22]D. J. L. Watt and A. H. Fabricius, "Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1 F2 plane," in *Leeds Working Papers in Linguistics and Phonetics* (2002), Vol. 9, pp. 159–173.