J. Yaghob (Ed.): ITAT 2015 pp. 194–199
Charles University in Prague, Prague, 2015

A New Method to Combine Probability Estimates from Pairwise Binary Classifiers

Ondrej Šuch[1], Štefan Beňuš[2], and Andrea Tinajová[3]

[1] University of Žilina and Slovak Academy of Sciences, Slovakia ondrejs@savbb.sk,
[2] Constantine the Philosopher University and Slovak Academy of Sciences, Slovakia sbenus@ukf.sk
[3] Slovak Academy of Sciences andrea.tinajova@gmail.com

Abstract: Estimating class membership probabilities is an important step in many automated speech recognition systems. Since binary classifiers are usually easier to train, one common approach to this problem is to construct pairwise binary classifiers. Pairwise models yield an over-determined system of equations for the class membership probabilities. Motivated by probabilistic arguments we propose a new way for estimating individual class membership probabilities, which reduces to solving a linear system of equations. A solution of this system is obtained by finding the unique non-zero eigenvector of total probability one, corresponding to eigenvalue one of a positive Markov matrix. This is a property shared by another algorithm previously proposed by Wu, Lin, and Weng. We compare properties of these methods in two settings: a theoretical three-way classification problem, and via classification of English monophthongs from TIMIT corpus. **Index Terms**: binary classifiers; multiclass classification; phoneme recognition; English vowels; TIMIT

1 Introduction

Probabilistic approach underlies most current automatic speech recognition (ASR) systems, and very likely also human speech perception. In many ASR systems a common task is to provide estimates of probabilities of a given sample belonging to multiple classes given the observed values of its features. These classes may represent various phonemes, diphones or other kinds of linguistic categories.

In machine learning it is easier to find the boundary between two classes rather than the boundary separating a class from many other classes [1]. Moreover, many discriminative models are naturally suited to pairwise classification, such as logistic regression, LDA or variants of SVM. Thus given k classes C_i, one can readily construct $\binom{k}{2}$ pairwise discriminative models. Let us denote by M_{ij} the model discriminating classes C_i and C_j. Suppose that M_{ij} is able not only to discriminate, but also to compute the pairwise class membership probability r_{ij} of an object X with features $\mathbf{f}$:

$$r_{ij} = r_{ij}(X) = p(X \in C_i | \mathbf{f}, X \in C_i \text{ or } X \in C_j). \quad (1)$$

Given the knowledge of $r_{ij}(X)$ the question is then to estimate multi-class probabilities p_i where

$$p_i = p_i(X) = p(X \in C_i | \mathbf{f}). \quad (2)$$

Inspired by Bradley-Terry model, Hastie and Tibshirani suggested [1] to require:

$$\frac{p_i}{p_i + p_j} = r_{ij} \quad (3)$$

$$\sum_i p_i = 1 \quad (4)$$

Note that there are $1 + \binom{k}{2}$ equations for k unknowns, so the system of equations is over-determined for $k \geq 3$ and it may be not possible to solve them.

In the next section we review several approaches which have been suggested to find approximate solution of (3). In Section 3 we will propose a new method to combine pairwise estimates. In Section 4 we will examine its performance with synthetic as well as real world acoustic data. In Conclusion we discuss findings of our experiments.

2 Existing Approaches

One natural requirement for an algorithm which determines probabilities p_i is that if the system (3) has a solution then the algorithm will find them exactly.

Several approaches satisfying this requirement are outlined in the work of Wu, Ling and Wen [2]. They consider the following functionals:

$$\delta_{HT} : \min_{\mathbf{p}} \sum_{i=1}^{k} [\sum_{j:j\neq i}^{k} (r_{ij}\frac{1}{k} - \frac{1}{2}p_i)]^2, \quad (5)$$

$$\delta_1 : \min_{\mathbf{p}} \sum_{i=1}^{k} [\sum_{j:j\neq i}^{k} (r_{ij}p_j - r_{ji}p_i)]^2, \quad (6)$$

$$\delta_2 : \min_{\mathbf{p}} \sum_{i=1}^{k} \sum_{j:j\neq i}^{k} (r_{ij}p_j - r_{ji}p_i)^2, \quad (7)$$

$$\delta_V : \min_{\mathbf{p}} \sum_{i=1}^{k} \sum_{j:j\neq i}^{k} (I_{\{r_{ij}>r_{ji}\}}p_j - I_{\{r_{ji}>r_{ij}\}}p_i)^2, \quad (8)$$

$$(9)$$

where I is the indicator function. Each of the four functionals is nonnegative. When the system (3) does have a solution, each functional is zero at, and only at the solution. One less satisfying feature of these approaches is that they lack probabilistic motivation, unlike the method we propose in the next section.

3 New Method

We will now describe our new algorithm. In general, one has $0 \leq r_{ij} \leq 1$. To avoid complications arising from degenerate cases we assume sharp inequalities $0 < r_{ij} < 1$, which poses no difficulty in practical applications.

Consider for a moment that an object X belongs to the class C_m. Then for judging its similarity to other classes one may restrict attention to the values r_{mj} (and $r_{jm} = 1 - r_{mj}$), since only classifiers M_{mj} were trained on values from the category C_m. But for those $k-1$ values equations (3) can be solved exactly, as we will now show.

We have

$$\sum_{j \neq m} \frac{1}{r_{mj}} = \sum_{j \neq m} \frac{p_m + p_j}{p_m} = (k-1) + \frac{1-p_m}{p_m}. \qquad (10)$$

This relation allows us to compute an estimate $p_m^{(m)}$ of p_m explicitly as

$$p_m^{(m)} = \left(\sum_{j \neq m} \frac{1}{r_{mj}} - (k-2) \right)^{-1}, \qquad (11)$$

where the upper index indicates that the estimate of p_m is computed by taking into account only values r_{mj}. The remaining probabilities can be then computed by the following formula:

$$p_j^{(m)} = p_m^{(m)} \cdot \left(\frac{1}{r_{mj}} - 1 \right). \qquad (12)$$

Now we repeat this argument for $m = 1, 2, \ldots, k$. In general the estimates of p_i thus obtained will be conflicting i.e. in general $p_j^{(m)} \neq p_j^{(n)}$, because given values r_{ij} may not allow for solving (3) consistently. We will now take inspiration from the probability law $p(A) = \sum_i p(A|B_i)p(B_i)$, if B_i is a partition of the probability space. We will require that the estimate $\hat{p}_i$ of p_i should satisfy the following linear system of equations:

$$\hat{p}_j = \sum_m p_j^{(m)} \hat{p}_m, \quad \text{for } j = 1, \ldots, k. \qquad (13)$$

These requirements can be interpreted as imposing self-consistency on the estimates $\hat{p}_i$. One readily checks that the matrix of the linear system (13) is Markov and positive, thus (13) has a one-dimensional space of solutions. Imposing an additional condition

$$\sum_m \hat{p}_m = 1 \qquad (14)$$

determines a unique estimate $\hat{p}_m$ of p_m.

4 Evaluation of the New Method

First note that our algorithm will yield the correct solution if the system (3) has a solution. In order to see that, one first checks using (10) and (11) that $p_m^{(m)} = p_m$ and $p_j^{(m)} = p_j$. It follows that the vector p_j satisfies equations (13) and (14). Since the solution of (13) and (14) is unique, the method will yield the correct solution. However, this is an ideal, very special situation that will generally not hold for $k \geq 3$.

We have opted to do comparison testing of the proposed method with the method of Wu, Ling and Wen [2] that minimizes functional δ_1 (6). The reason is that that method also involves the construction of a positive Markov matrix whose solution is their estimate of p_m. We conduct two experiments: one is an artificial three-way classification problem, and the other a vowel recognition task.

4.1 Three-Way Classification

The system of equations (3) becomes over-determined for $k = 3$. If one of the classifiers is unreliable then the system (3) will not have a solution. In this section we present the results of a synthetic experiment for three-way classification.

In our experiment we assume that only classifier M_{23} is unreliable. In other words we assume that classifiers M_{12} and M_{13} discriminating respectively categories C_1 versus C_2 and C_1 versus C_3 yield precise estimates of r_{12} and r_{13}. For a fixed value p_1, p_2 we thus set $r_{12} = p_1/(p_1 + p_2)$ and $r_{13} = p_1/(p_1 + p_3) = p_1/(1 - p_2)$. Let $\hat{p}_m$ and p_m^{Wu} denote our and Wu's estimates of p_m. As r_{23} varies in interval $(0,1)$, define the absolute errors

$$\Delta = \sup_{i, r_{23}} |\hat{p}_i - p_i|, \qquad (15)$$

$$\Delta_{\mathrm{Wu}} = \sup_{i, r_{23}} |p_i^{\mathrm{Wu}} - p_i|, \qquad (16)$$

and the relative error

$$\Delta_{\mathrm{Wu}}^{\mathrm{rel}} = \sup_{i, r_{23}} |p_i^{\mathrm{Wu}} - \hat{p}_i|. \qquad (17)$$

The results of our experiment are shown in Table 1. From the table it is clear that sometimes our method gives more precise estimates, but for other values of p_1, p_2, Wu's method will yield more precise results. However, in all cases, the relative error between our results and Wu's results is smaller than the absolute errors, often by an order of one magnitude.

4.2 Vowel Recognition

Unlike consonants, vowels may be perceived non-categorically by listeners [3], making it a good testing ground for multi-class probabilistic estimates. We opted for English language, because it has a large variety of vowels and because there are large corpora of annotated speech available. We worked with TIMIT, a phonetically segmented corpus of American English [4]. Our categories consisted of 15 monophthongs as shown in Table 2. For

p_1	p_2	Δ	Δ_{Wu}	Δ_{Wu}^{rel}
0.05	0.05	0.66	0.7	0.09
0.1	0.1	0.57	0.61	0.09
0.85	0.1	0.07	0.05	0.05
0.85	0.05	0.07	0.05	0.05
0.05	0.85	0.66	0.70	0.1
0.1	0.85	0.58	0.61	0.06
0.33	0.33	0.21	0.22	0.05

Table 1: Errors of estimation for various values of p_1 and p_2

vowel	sample word	sample word's transcription
iy	beet	bcl b IY tcl t
ih	bit	bcl b IH tcl t
eh	bet	bcl b EH tcl t
ae	bat	bcl b AE tcl t
aa	bott	bcl b AA tcl t
ah	but	bcl b AH tcl t
ao	bought	bcl b AO tcl t
uh	book	bcl b UH kcl k
uw	boot	bcl b UW tcl t
ux	toot	tcl t UX tcl t
er	bird	bcl b ER dcl d
ax	about	AX bcl b aw tcl t
ix	debit	dcl d eh bcl b IX tcl t
axr	butter	bcl b ah dx AXR
ax-h	suspect	s AX-H s pcl p eh kcl k tcl t

Table 2: Sample words containing 15 different monophong sounds of American English as segmented in TIMIT corpus

each of the categories we randomly chose their realizations from the set of male speakers in the corpus. Each realization was analyzed with a window 512 samples wide (at 16kHz sampling rate its length was 32ms). If the center of the window was less than 256 samples away from the next phoneme, it was proportionally less likely to be selected into our dataset. We have trained pairwise classifiers using linear discriminant analysis (LDA). The feature set was log-periodogram, where the analysis window was weighted with Hanning window before computing FFT.

We have performed comparison testing of our and Wu's method by selecting 500 random samples from the test subset. Per phone results are shown in Table 3. The key statistics is that overall there was 96% agreement between most-likely classifications by our method and Wu's method.

The overall success rate was slightly below 40% for both our and Wu's method. Due to the limitations of the features (no F0, no vowel duration, no dynamic information, no multiframe data), suboptimal performance may be expected. For instance without intensity baseline, it is nearly impossible to correctly distinguish some accented vowels.

vowel	success rate	Wu's success rate	agreement
iy	48 %	48 %	96.6%
ih	21 %	21 %	94.8 %
eh	22 %	23 %	95.4 %
ae	60 %	60 %	94.4 %
aa	48 %	48 %	96.2 %
ah	20 %	21 %	94.6 %
ao	60 %	61 %	97.2 %
uh	18 %	18 %	95 %
uw	40 %	39 %	96.4 %
ux	40 %	40 %	97.4 %
er	34 %	35 %	95.6 %
ax	31 %	31 %	96.4 %
ix	16 %	18 %	94.4 %
axr	48%	46 %	96.2 %
ax-h	81 %	81 %	98.8 %

Table 3: Evaluation of our and Wu's [2] methods on individual monophthongs from the test data from TIMIT corpus. The first column indicates agreement between classification by our method and TIMIT annotation, the second column the statistics for method of Wu et al, and the third column indicates how often our method and Wu's method agreed on the most-likely classified class.

We decided to do a more detailed case study. From the test subset we have chosen sentence SA1 spoken by speaker MREB0 and examined each monophthong at two points in time. The first was 5 milliseconds after the onset, and the other one approximately near the vowel's center. The results are shown in Table 4.

Likelihoods of most likely estimates of our and Wu's method are again quite close. There are two differences between onset and center predictions. The first one is misprediction of /er/ at the beginning of the word 'greasy', which is quite understandable, since the vowel is preceded by /r/. To gain an insight into the other mispredictions as well as deeper insight into dynamical behavior of the resulting multiclass classifier we present time plots in Fig. 1. In Fig. 1a the mis-classification of /iy/ instead of TIMIT's /ix/ in the word 'in' is shown. We speculate that the problem might be attributed to greater weight put on F2, that is relatively high and within the region for /iy/, compared to F1 that is quite high and definitely within the region for /ix/. In other words, the vowel might be a bit fronter than canonical /ix/. In Fig. 1b, the first vowel of 'greasy' is mis-classified as /ux/ instead of TIMIT's /iy/.

This problem might be attributed to coarticulation from the flanking consonants. The first vowel does have lower F2, which is plausibly responsible for /ux/ prediction, but it is preceded by /r/, which is commonly associated with lip protrusion, which lowers F2. In Fig. 1c in the vowel of word 'wash', we see that it is only in the beginning in the word 'wash' that the classifier gives more weight to /ao/, and then it increasingly agrees that the vowel is /aa/.

offset	TIMIT label	Wu's method		our method	
3831	iy	**iy**	80.1 %	**iy**	79.9 %
6053	ae	**ae**	79.7 %	**ae**	79.6 %
9187	axr	**axr**	62.2 %	**axr**	61.6 %
11780	aa	**aa**	32.9 %	**aa**	32.6 %
19677	ux	**ux**	60.3 %	**ux**	58.2 %
25544	ix	iy	66.4 %	iy	64.9 %
28905	iy	er	41.8 %	er	40.3 %
31328	iy	**iy**	53.4 %	**iy**	53.3 %
34210	aa	ao	76.3 %	ao	75.8 %
39080	ao	aa	77.1 %	aa	76.9 %
40680	er	axr	56.8 %	axr	56.3 %
42512	ao	**ao**	87.2 %	**ao**	87.1 %
46827	ih	iy	58.3 %	iy	57.9 %
48248	axr	**axr**	52.1 %	**axr**	52.4 %

(a) 5ms after vowel's start

offset	TIMIT label	Wu's method		our method	
4200	iy	**iy**	83.2 %	**iy**	82.9 %
6800	ae	**ae**	83.7 %	**ae**	83.6 %
9600	axr	**axr**	50.6 %	**axr**	50.7 %
12500	aa	**aa**	80.7 %	**aa**	79.5 %
21000	ux	**ux**	67.2 %	**ux**	66.4 %
25800	ix	iy	55.5 %	iy	53.3 %
29000	iy	ux	23.5 %	ux	22.8 %
31800	iy	**iy**	72.8 %	**iy**	72.7 %
35000	aa	**aa**	57.6 %	**aa**	57.7 %
39600	ao	aa	78.8 %	aa	78.5 %
41500	er	axr	66.4 %	axr	66.4 %
43500	ao	**ao**	86.3 %	**ao**	86.3 %
47500	ih	ux	37.9 %	ux	37 %
49000	axr	**axr**	71.1 %	**axr**	71.1 %

(b) near the center of the vowel

Table 4: Results of monophthong classification using spectral information in 32ms window centered at the offset indicated in the first column. Vowels were extracted from sentence SA1 spoken by speaker MREB0 from region 1 (New England). Most likely classes are shown computed by Wu's method and our method together with multi-class likelihoods.

In this particular case, we conclude that our classification is closer to the phonetic realization than TIMIT's. The beginning of the vowel is influenced by the preceding /w/ with lip rounding similar to /ao/. The rest of the vowel sounds like an /aa/ to phonetically trained listeners, and the formant values correspond to this perception. Finally, Fig. 1d shows the preference for /aa/ as the first vowel of 'water' in our model over /ao/ in TIMIT's. Similarly to Fig. 1c, this vowel sounds more, and its formant values correspond to our model more, than to TIMIT's. It should be noted, however, that /ao/ and /aa/ have merged in several American dialects and more tokens would be needed for a more thorough analysis.

A common way to improve the performance in automatic speech recognition is to tune the parameters of the system for a particular speaker. To that end we carried one more experiment. We extracted formants for TIMIT vowels spoken by speaker MREB0 using package *phonTools* in R [5]. Next we performed pairwise LDA training as previously but this time used values F1 and F2 for features rather than the log-periodogram. These first two formants are key perceptual features of vowels [6, 7, 8, 9]. Finally, we performed multiclass classification on the first vowel in the word 'water'. The formants contours for this vowel are shown in Fig. 2.

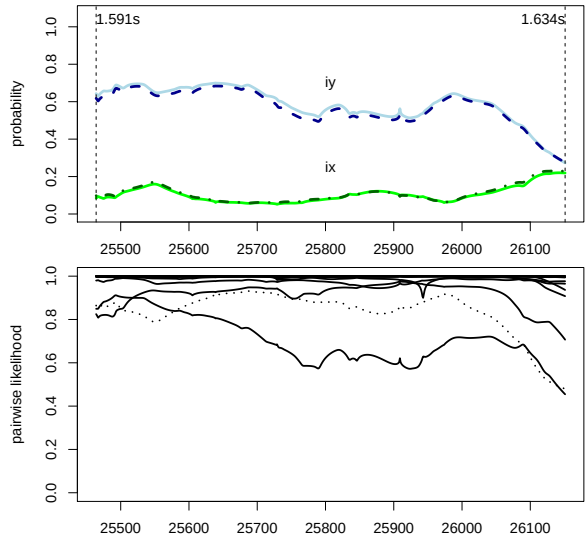
The somewhat suprising results are shown in Fig. 3. One would expect that it would have little problem with classification of the vowel. As seen in Fig. 3, except for a brief start, the classifier overwhelmingly believes that the phoneme is much closer to /aa/ than TIMIT annotated /ao/. However, compared to Fig. 1d the likelihood of /aa/ is markedly smaller near the vowel's boundaries.

5 Conclusions

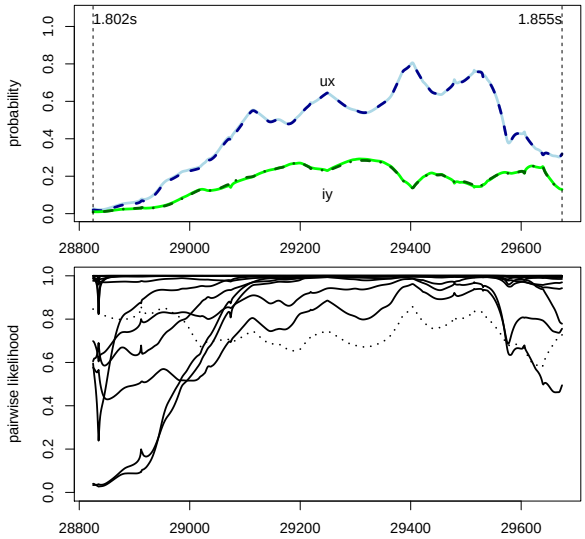
We have described a new method for combining probability estimates from pairwise classifiers. It is quite general and for its application needs only pairwise classifiers that provide posterior likelihoods. We believe that since the rationale for our method is probabilistically motivated, it has the potential to edge out other methods in practice. In particular by its construction it avoids the problem of 'pairwise coupling' approaches pointed out by G. Hinton [1, pg. 467]. Another important feature is that the resulting probabilities are computed as the dominant eigenvector of a Markov matrix, allowing for efficient computation via iterations when the matrix of binary likelihoods varies slowly in time. Finally, since the method is not hierarchical, it avoids compounding of errors common in hierarchical approaches.

In presented synthetic and phonetic experiments its performance was very close to a method previously suggested by Wu [2]. The classification of English vowels was suboptimal, but that may not be indicative of performance in real world scenarios for several reasons.

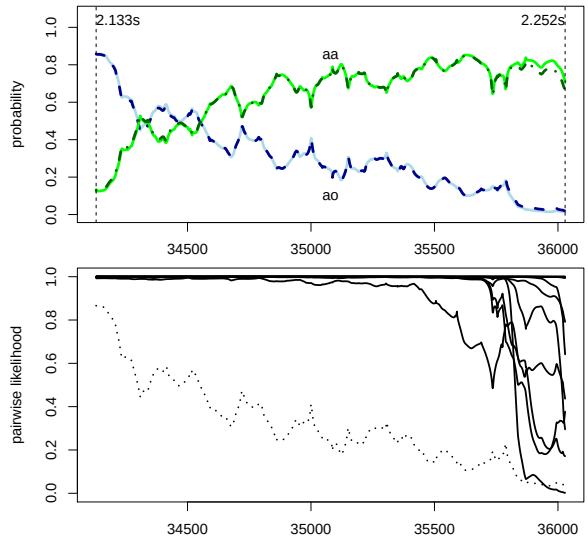
- We have used all TIMIT vowel categories, some of which are in previously published performance benchmark tests fused because they are extremely hard to discriminate.
- Other pairwise classifiers, for instance logistic regression or SVM may yield better results.
- Based on the last experiment presented, we question whether TIMIT annotation is consistent throughout the corpus even for individual speakers.



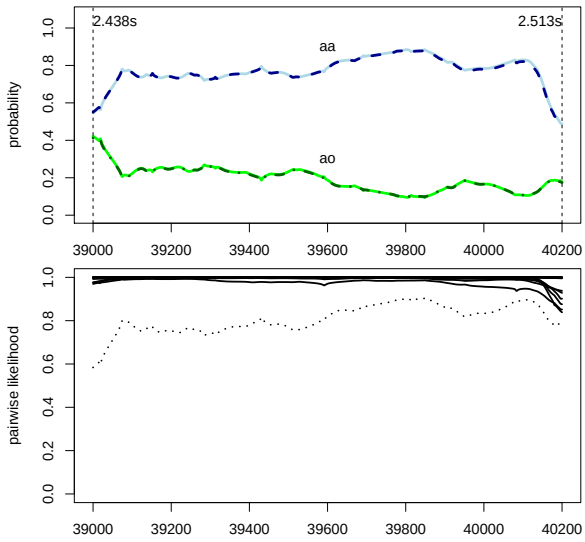
(a) TIMIT annotation is /ix/ in the word ‘in’. We considered an alternative classification that the vowel is /iy/.



(b) TIMIT annotation is /iy/ for the first vowel in the word ‘greasy’. We considered an alternative classification that the vowel is /ux/.



(c) TIMIT annotation is /aa/ in the word ‘wash’. We considered an alternative classification that the vowel is /ao/.



(d) TIMIT annotation is /ao/ for the first vowel in the word ‘water’. We considered an alternative classification that the vowel is /aa/.

Figure 1: Time series plots of multiclass and pairwise classification likelihoods for four vowels in sentence SA1 spoken by MREB0. The top plot in each subfigure shows multiclass likelihoods, and the bottom plot shows binary classification likelihoods r_{ij}. In multiclass plots, dashed dark curve indicates the likelihood of the alternative hypothesis and dark dash-dotted curve that of TIMIT annotation computed by our method (i.e. $\hat{p}_i$). Solid curves in multiclass plots indicate corresponding but visually nearly indistinguishable estimates obtained via Wu’s method. In binary plots we plot likelihoods of the alternative hypothesis against all other classes. The dotted curve in each binary plot indicates likelihood of the alternative hypothesis compared to the TIMIT annotation.

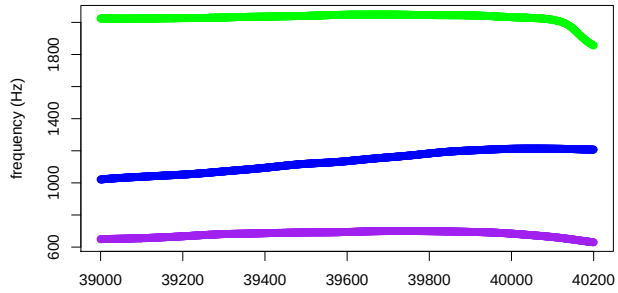


Figure 2: Formant contours F1-F3 for the first vowel of word 'water' in sentence SA1 spoken by MREB0.

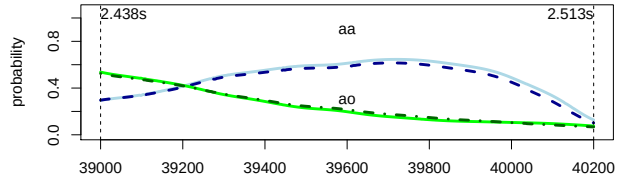


Figure 3: Time series plots of multiclass likelihoods for the first vowel in the word 'water' spoken in sentence SA1 by speaker MREB0. Dark dashed curve indicates likelihood of /aa/, whereas dot-dashed curve indicates likelihood of /ao/. Solid curves, as in Fig. 1, indicate estimate by Wu's method.

Further experiments with a complete ASR system may shed more light on the applicability of the proposed algorithm.

Acknowledgements

Our research was supported by the project University Science Park ITMS 26220220184 and grants APVV-0219-12, APVV-14-0560 and VEGA 2/0197/15. The authors are thankful to Paul Foulkes, K. Bachratá, and Martin Klimo for helpful discussion.

References

[1] Hastie, T. H., Tibshirani, R.: Classification by pairwise coupling. Annals of Statistics **26 (2)** (1998), 451–471

[2] Wu, T. -F., Lin, C. -J., Weng, R.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research **5** (2004), 975–1005

[3] Fry, D., Abramson, A., Eimas, P., Liberman, A.: The identification and discrimination of synthetic vowels. Language and Speech **5** (1962), 171–189

[4] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., Zue, V.: TIMIT acoustic-phonetic continuous speech corpus, [Online], 1993.

[5] Barreda, S.: phonTools: functions for phonetics in R, R package version 0.2-2.0, 2014

[6] Potter, R., Steinberg, J.: Toward the specification of speech. J. Acoust. Soc. Amer. **22 (6)** (1950), 807–820

[7] Peterson, G., Barney, H.: Control methods used in a study of vowels. J. Acoust. Soc. Amer. **24 (2)** (1952), 175–184

[8] Turner, R., Patterson, R.: An analysis of the size information in classical formant data: Peterson and Barney (1952) revisited. J. Acoust. Soc. Jpn. **33** (2003)

[9] Kiefte, M., Nearey, T., Assmann, P.: Vowel perception in normal speakers. In: Handbook of vowels and vowel disorders, M. Ball and F. Gibbon, (Eds.) Psychology Press, 2012, ch. 6, 160–185