# COMPARISONS OF SPEAKER RECOGNITION STRENGTHS USING SUPRASEGMENTAL DURATION AND INTENSITY VARIABILITY: AN ARTIFICIAL NEURAL NETWORKS APPROACH

Lei He[1], Ulrike Glavitsch[2], Volker Dellwo[1]

[1] Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Switzerland
[2] EMPA, Swiss Federal Laboratories for Materials Science and Technology, Dübendorf, Switzerland
lei.he@uzh.ch; ulrike.glavitsch@empa.ch; volker.dellwo@uzh.ch

## ABSTRACT

This study compares the speaker recognition strengths based on suprasegmental duration and intensity variability in the speech signal using artificial neural networks. Such algorithm can well capture the nonlinear effects in the data, and is more robust against noise in the data. Three rounds of classification tasks were performed with 1) duration metrics, 2) intensity metrics, and 3) the combination of duration and intensity metrics as the independent variables. The results indicated that both intensity and combined metrics significantly outperformed the duration metrics. Moreover, the combination of intensity and duration metrics showed higher probability of improved speaker classifications than intensity metrics over duration metrics.

**Keywords**: duration variability, intensity variability, speaker recognition, artificial neural networks

## 1. INTRODUCTION

Speech production is a complicated process underpinned by sophisticated neuromuscular programming for the motor control of speech organs [6]. The movements of speech organs, like the movements of other parts of the body (see [20, 28] for the example of human gait), are highly idiosyncratic, and such idiosyncrasy should find its acoustic correlates in the speech signal, particularly in the time domain.

In a previous research project of our laboratory [7, 17], the researchers applied the widely used rhythm (duration) metrics (such as ΔC, ΔV, %V, varcoC, varcoV, rPVI-C, and nPVI-V), which were originally developed by [5, 9, 19, 22, 27] to segregate traditionally categorised “stress-” and “syllable-timed” languages [1, 4, 18, 21], and have found significant speaker individualistic temporal characteristics [7, 17].

Along the same line of reasoning, we also hypothesised that individualistic movements as well as anatomical peculiarities of speech organs should result in idiosyncratic energy distribution in the speech signal, and quantifying intensity variability in the signal should capture such idiosyncrasy. Enlightened by the duration metrics, we [2, 10] have developed intensity metrics (please see the appendix) to calculate the syllabic intensity variability (either the mean RMS or peak RMS of each syllable), and the results showed that significant effects of the speaker existed for all the intensity metrics [2, 10].

Our long-term research goal is to explore how successful automatically extracted temporal as well as intensity features will contribute to speaker recognitions, so that they can be implemented in real speaker recognition systems. The present study is an intermediate step towards this goal: we used the human labelled TEVOID corpus (see [7, 17] and §2.1 for more information) and calculated the duration and intensity metrics which were fed into the well-established classification algorithm of artificial neural networks (abbreviated as ANN hereafter), and found that a combination of both duration and intensity metrics gave the best performance of offline speaker recognitions.

The reasons for choosing ANNs were threefold: 1) nonlinear effects in the data, which cannot be controlled for *a priori*, can be modelled by the algorithm [12]; 2) being an eager learner, the ANN generalises the training data before receiving queries from the test data [25], so that the classification is less susceptible to noise; and 3) as a commonly accepted classification algorithm, it can be used as a reference of success for developing new algorithms, which is also in our research pipeline. Primers to the ANN are available as [8, 14, 16], and phonetic research using ANNs include [15, 23, 24, 26], where the latter two focus on speaker recognition.

## 2. METHOD

### 2.1. The Corpus

The TEVOID (*Te*mporal *Vo*ice *Id*iosyncrasy) corpus [7, 17] was constructed to investigate speaker individualistic temporal characteristics in the speech signal. For the present study, the read speech of the corpus was analysed (16 native speakers of Zürich German × 256 sentences = 4,096 sentences; wav audio format; sampling frequency = 44.1 kHz; quantisation depth = 16 bits). For more of the corpus construction, please refer to [7, 17].

### 2.2. Measurements

All the sound files in the corpus were labelled using Praat [3]. Tiers containing on- and off-sets of vocalic and consonantal intervals were employed for the calculations of %V, varcoV, nPVI-V, varcoC, and nPVI-C. Tiers containing on- and off-sets of voiced intervals were used for the calculations of %VO, varcoVO, and nPVI-VO. Tiers containing syllable boundaries as well as syllable peaks were applied to compute varcoPeak, nPVI-Peak, stdevM, varcoM, rPVIm, nPVIm, stdevP, varcoP, rPVIp and nPVIp. Descriptions of all the measures are listed in the Appendix. Praat scripts were applied for the computations, and the results were saved as tab-delineated files before exporting to SPSS [13] for the constructions of neural networks (multilayers perceptron).

### 2.3. ANN Topologies

The corpus was randomly partitioned into a training set (70% of the corpus) and a test set (30% of the corpus). Three ANNs were modelled based on the same partitioned corpus using *a*. duration metrics only, *b*. intensity metrics only, and *c*. duration cum intensity metrics. The choices of ANN typologies were the same for all three models except the input covariates, which were the duration, intensity and combined metrics respectively. Table 1 presents more details of the ANN architectures, which were configured on a semi-arbitrary basis, because the purpose of the study was to compare the classification strengths rather than maximising classification rates. Nonetheless, we did venture a more complicated configuration of the networks (two hidden layers with 100 neurons in each), but the recognition time increased dramatically without remarkable improvements of the recognition rates.

## 3. RESULTS AND DISCUSSION

### 3.1. Speaker Recognition Rates

The average speaker recognition rates yielded from the ANNs in the training set were 17.3% (duration only), 33.1% (intensity only), and 42.3% (duration cum intensity). The mean recognition rates calculated from the test set were 14.2% (duration only), 30.3% (intensity only), and 36.9% (duration cum intensity). Table 2 shows more descriptive statistics of speaker recognition rates in different choices of metrics. Figures 1 and 2 present the breakdowns of classification rates for each speaker in both training and test sets.

**Table 1**: ANNs fitting information.

| Input |
|---|
| Input covariates: duration metrics only; intensity metrics only; duration cum intensity metrics<br>Rescaling method for covariates: Standardised |
| Hidden layer (1 hidden layer) |
| Number of neurons in the hidden layer: 10 + 1 bias<br>Activation function: Sigmoid |
| Output |
| Dependent variable: speaker<br>Activation function: Softmax<br>Error function: Cross-entropy<br>------------------------------------------------------------<br>NB All networks are feedforward without recursions. |

**Table 2**: Descriptive statistics of speaker recognition rates (in %) with different independent variables.

| | mean | std. dev. | std. err. | min. | max. |
|---|---|---|---|---|---|
| (i) Training set | | | | | |
| D* | 17.3 | 11.6 | 2.9 | 3.7 | 38.4 |
| I* | 33.1 | 19.5 | 4.9 | 10.1 | 77.6 |
| C* | 42.3 | 17.3 | 4.3 | 14.8 | 77.6 |
| (ii) Test set | | | | | |
| D* | 14.2 | 13.1 | 3.3 | 0.0 | 39.4 |
| I* | 30.3 | 19.2 | 4.8 | 9.5 | 73.2 |
| C* | 36.9 | 19.5 | 4.9 | 11.9 | 67.1 |

* D = duration metrics; I = intensity metrics; C = duration cum intensity metrics.

**Figure 1**: Speaker recognition rates in the training set (X-axis: speaker ID; Y-axis: recognition rate in %). The horizontal dashed line indicates the chance level (100% ÷ 16 ≅ 6.3%).

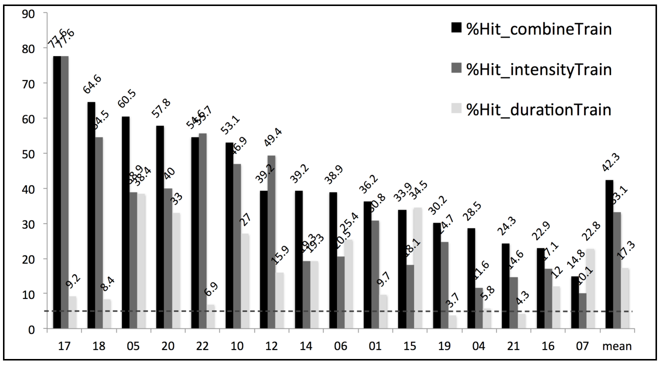
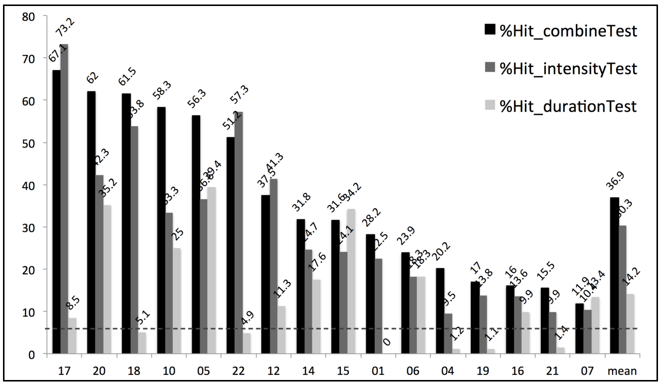


**Figure 2**: Speaker recognition rates in the test set (X-axis: speaker ID; Y-axis: recognition rate in %). The horizontal dashed line indicates the chance level (100% ÷ 16 ≅ 6.3%).



### 3.2. Comparisons of Recognition Strengths

First of all, the distribution normalities of the recognition rates from three ANN models (both training and test data) were evaluated using the Shapiro-Wilk test, and the results indicated no serious deviations from normality (all $p$ values ≥ 0.05).

**Table 3**: Results of Bartlett's tests and paired *t*-tests (2-sided).

| Train vs. | Bartlett's tests | | *t*-tests | |
|---|---|---|---|---|
| Test | $K^2$ ($df$=1) | $p$ | $t$ ($df$=15) | $p$ |
| duration | 0.2107 | >0.6 | 3.7206 | =0.002 |
| intensity | 0.0056 | >0.9 | 2.0576 | >0.05 |
| combined | 0.1908 | >0.6 | 3.9097 | =0.001 |

**Table 4**: Results of Bartlett's tests and ANOVAs.

| | Bartlett's tests | | ANOVAs | |
|---|---|---|---|---|
| | $K^2$ ($df$=2) | $p$ | $F$ ($df$=2,45) | $p$ |
| durationTrain<br>intensityTrain<br>combineTrain | 3.9451 | >0.1 | 9.4043 | <0.0004 |
| durationTest<br>intensityTest<br>combineTest | 2.6837 | >0.2 | 7.1583 | <0.002 |

Paired samples *t*-tests were run in order to compare if the training set recognitions were significantly better than the test set recognitions. Bartlett's tests indicated the data variances were homogenous; therefore, no adjustments were needed. Tables 3 shows the statistical results: only the recognition rates between training and test sets using intensity measures were not significantly different. The results indicated some degrees of over-adaptations of the training data, which is one of the weaknesses of the ANN [16].

Finally, univariate ANOVAs were utilised and the results indicated that significant effects of the metrics choice existed (Table 4 shows the statistics). Bartlett's tests confirmed the equalities of variances, so no adjustments were necessary (also see Table 4).

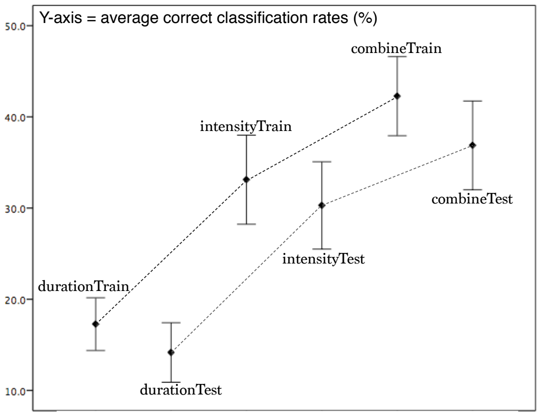
Post hoc pairwise comparisons (Bonferroni adjusted) indicated that in the training set, intensity metrics and intensity cum duration metrics were significantly better than duration metrics alone at idendifying speakers ($^{\text{Train}}p_{\text{intensity:duration}} < 0.03$, $^{\text{Train}}p_{\text{combine:duration}} < 0.0003$). However, the intensity metrics and the combined metrics were not significantly different ($^{\text{Train}}p_{\text{intensity:combine}} > 0.4$). The test set showed similar patterns: intensity metrics and combined metrics performed significantly better in speaker recognitions, but the intensity metrics and combined metrics were not significantly different ($^{\text{Test}}p_{\text{intensity:duration}} < 0.04$, $^{\text{Test}}p_{\text{combine:duration}} < 0.002$, $^{\text{Test}}p_{\text{intensity:combine}} > 0.8$). Figure 3 visualises the patterns.

This suggests that although both duration and intensity measures had significant speaker effects [7, 17, 10, 2], intensity variability in the speech signal showed more strength compared with duration metrics to classify speakers. However, albeit the speaker discriminability of duration cum intensity measures were not significantly different from intensity measures in post hoc tests, the significance level of $p_{\text{intensity:duration}}$ (0.03) was a hundred folds the significance level of $p_{\text{combine:duration}}$ (0.0003) in the training set. In the test set, the significance level of $p_{\text{intensity:duration}}$ (0.04) was twenty folds the significance level of $p_{\text{combine:duration}}$ (0.002).

In other words, alghough both combined metrics and intensity metrics were significantly better than duration metrics for the recognition of speakers in both training and test sets, the probability that the combined metrics significantly improved over duration metrics increased 2.97 percentage points than the intensity metrics alone for the training data ($(1 - {}^{\text{Train}}p_{\text{combine:duration}}) - (1 - {}^{\text{Train}}p_{\text{intensity:duration}}) = (1 - 0.0003) - (1 - 0.03) = 0.0297$), and 3.8 percentage points for the test data ($(1 - {}^{\text{Test}}p_{\text{combine:duration}}) - (1 - {}^{\text{Test}}p_{\text{intensity:duration}}) = (1 - 0.002) - (1 - 0.04) = 0.038$).

In addition, it was also assumed that only two rounds of speaker classification tasks (on the basis of intensity cum duration metrics and intensity metrics alone) had been performed. Paired samples *t*-tests showed that the combined metrics significantly improved recognition rates than intensity metrics alone in the training set ($t$ = 4.1527, 2-sided, $p < 0.0009$, with $df = 15$) and test set ($t = 2.9588$, 2-sided, $p < 0.01$, with $df = 15$).

**Figure 3**: Error bar graph showing general speaker recognition rates (mean ± 1 standard error) with different metrics choices.



## 5. CONCLUSION

The present study explored speaker recognition strengths using duration variability, intensity variabily and the two combined in the speech signals with the feedforward ANN. The results suggested that intensity metrics and intensity cum duration metrics were stronger in speaker recognitions.

Compared with our previous studies, we can see that speaker recognition success depends on the recognition algorithms as well. For instance, the TEVOID corpus with the intensity metrics as described in the current study alone yielded different degrees of correct classifications using the *k*-nearest neighbours (*k*NN), feedforward ANN, and multinomial logistic regressions [2, 11], where the *k*NN showed poorest performance (average hit rate ≅ 12%), and the logistic regression showed the best performance (average hit rate ≅ 38%). There is potential to design or optimise recognition algorithms and achieve higher recognition rates with the combination of intensity and duration measures.

In addition, we have also observed that although significant between-speakers variability has been proven by statistical tests, it does not necessarily entail high recognition rates using available classification algorithms (duration metrics in particular).

Open questions for future research are how robust the presented measures are in the context of degraded and distorted speech. Also, how speaker recognition rates would increase if the duration and intensity measures are coupled with spectral measurements is worth further examinations. Moreover, how should the metrics and classification algorithms be optimised is also subjective to further investigations. On the engineering side, a fully automatic extraction of the metrics from the acoustic signal is also in our research agenda.

## 6. APPENDIX−METRICS DESCRIPTIONS

### 6.1. Duration Metrics

- %V: Percentage of vocalic interval durations out of the total sentential duration.
- %VO: Percentage of voiced interval durations out of the total sentential duration.
- varcoC: Variation coefficient (standard deviation ÷ mean) of consonantal interval durations.
- nPVI-C: Mean of locally averaged pairwise consonantal interval duration differences.
- varcoV: Variation coefficient of vocalic interval durations.
- nPVI-V: Mean of the locally averaged pairwise vocalic interval duration differences.
- varcoPeak: Variation coefficient of syllabic peak-to-peak interval durations.
- nPVI-Peak: Mean of the locally averaged pairwise syllabic peak-to-peak interval duration differences.
- varcoVO: Variation coefficient of voiced interval durations.
- nPVI-VO: Mean of the locally averaged pairwise voiced interval duration differences.

### 6.2. Intensity Metrics

- stdevM/P: Standard deviation of mean/peak intensity of each syllable.
- varcoM/P: Variation coefficient of mean/peak intensity of each syllable.
- rPVIm/p: Mean of the pairwise mean/peak intensity differences of consecutive syllables.
- nPVIm/p: Mean of the locally averaged pairwise mean/peak intensity differences of consecutive syllables.

Mathematical formulae of these metrics can be found in [17] and [10, 11].

## 7. ACKNOWLEDGEMENTS

This study is supported by the Gebert Rüf Stiftung (Grant No. GRS-027/13) and the Swiss National Science Foundation (Grant No. 100015_135287). The authors would like to thank Adrian Leemann and Marie-José Kolly for their work on the TEVOID corpus

## 8. REFERENCES

[1] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

[2] He, L., Dellwo, V. submitted. The role of syllable intensity in between-speaker rhythmic variability.

[3] Boersma, P., Weenink, D. 2014. Praat: doing phonetics by computer (version 5.3.65). http://www.praat.org/.

[4] Classe, A. 1939. *The Rhythm of English Prose* Oxford: Blackwell.

[5] Dellwo, V. 2006. Rhythm and speech rate: A variation coefficient for deltaC. In: Karnowski, P., Szigeti, I. (eds), *Language and Language Processing*, Frankfurt: Peter Lang, 231-241.

[6] Dellwo, V., Huckvale, M., Ashby, M. 2007. How is individuality expressed in voice? An introduction to speech production and description for speaker classification. In: Müller, C., (ed), *Speaker Classification I*. Berlin and Heidelberg: Springer Verlag, 1-20.

[7] Dellwo, V., Leemann, A., Kolly, M.-J. 2012. Speaker idiosyncratic rhythmic features in the speech signal. *Proc. Interspeech 2012* Portland, 1582-1585.

[8] Gershenson, C. 2003. Artificial neural networks for beginners. arXiv: cs/0308031. http://arxiv.org/pdf/cs/0308031.pdf.

[9] Grabe, E., Low, E. L. 2002. Durational variability in speech and rhythm class hypothesis. In: Warner, N., Gussenhoven, C. (eds), *Papers in Laboratory Phonology 7*. Berlin: Mouton de Gruyter, 515-543.

[10] He, L., Dellwo, V. 2014. Speaker idiosyncratic variability of intensity across syllables, *Proc. Interspeech 2014*, Singapore, 233-237.

[11] He, L., Glavitsch, U., Dellwo, V. 2014. Automatic speaker identification using syllable intensity variability: an initial attempt using the *k*NN classifier. Abstract presented at Phonetik & Phonologie 10, Konstanz. http://ling.unikonstanz.de/pages/conferences/pp10/abstracts/He_pp10.pdf.

[12] Heaton, J. 2011. *Introduction to the Math of Neural Networks (Beta-1)*. Chesterfield: Heaton Research Inc. (http://www.heatonresearch.com).

[13] IBM Corp. 2013. IBM SPSS Statistics for Macintosh (version 22.0). Armonk, NY: IBM Corp.

[14] Jain, A. K., Mao, J. 1996. Artificial neural networks: a tutorial. *Computer*. 29, 31-44.

[15] Jassem, W., Grygiel, W. 2004. Off-line classification of Polish vowel spectra using artificial neural networks. *J. IPA*. 34, 37-52.

[16] Krogh, A. 2008. What are artificial neural networks? *Nat. Biotechnol*. 26, 195-197.

[17] Leemann, A., Kolly, M.-J., Dellwo, V. 2014. Speech-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Sci. Int*. 238, 59-67.

[18] Lloyd James, A. 1940. *Speech Signals in Telephony* London: Sir Isaac Pitman & Sons.

[19] Low, E. L., Grabe, E., Nolan, F. 2000. Quantitative characterization of speech rhythm: Syllable-timing in Singapore English. *Lang. Speech*. 43, 377-401.

[20] Matovski, D. S., Nixon, M. S., Mahmoodi, S., Carter, J. M. 2010. The effect of time on the performance of gait biometrics. In: *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*. Washington DC. http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5618724.

[21] Pike, K. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

[22] Ramus, F., Nespor, M., Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*. 73, 265-292.

[23] Sałapa, K., Trawińska, A., Roterman-Konieczna, I. 2013. Forensic voice comparison by means of artificial neural networks. *Bio-Alg. Med-Syst*. 9, 191-197.

[24] Sałapa, K., Trawińska, A., Roterman, I., Tadeusiewicz, R. 2014. Speaker identification based on artificial neural networks. Case study: the Polish vowel a (pilot study). *Bio-Alg. Med-Syst*. 10, 91-99.

[25] Söder, O. 2008. kNN classifiers 1: What is a kNN classifier? http://www.fon.hum.uva.nl/praat/manual/kNN_classifiers_1__What_is_a_kNN_classifier_.html

[26] Weenink, D. 2006. Speaker-Adaptive Vowel Identification. Doctoral dissertation, Universiteit van Amsterdam.

[27] White, L., Mattys, L. S. 2007. Calibrating rhythm: first language and second language studies. *J. Phonet*. 35, 501-522.

[28] Yoo, J.-H., Hwang, D., Moon, K.-Y., Nixon, M. S. 2008. Automated human recognition by gait using neural network. In: *First Workshops on Image Processing Theory, Tools and Applications*. Sousse. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4743792.