Robustness of Spectral Moments: a Study using Voice Imitations

Erik J. Eriksson[*], **Luis F. Cepeda**[†], **Robert D. Rodman**[†],
Kirk P. H. Sullivan[*], **David F. McAllister**[†], **Donald Bitzer**[†] & **Pam Arroway**[‡]

[*]Department of Philosophy and Linguistics, Umeå University, Sweden
[†]Department of Computer Science, North Carolina State University, USA
[‡]Department of Statistics, North Carolina State University, USA

Abstract

The accuracy of forensic identification / verification methods are constantly under question. How accepted an expert witness's opinion is in a court room setting is highly dependent on both his, or her reputation as an expert and the method he or she used to perform the analysis presented in court. The method used must show robustness and accuracy to a known degree of confidence. This is equally true for forensic speaker identification / verification tasks. For such a task, a method using spectral moments has been devised at North Carolina State University, USA. By using imitations that confuse humans the robustness of the system has been tested. Five imitations, of a famous Swedish politician of which two were professional imitations, and a set of foils, were run through the method. Although the results showed that the method has limitations, confusion analysis demonstrated that the method was insensitive to voice imitation.

1. Introduction

Automatic speaker identification for verification purposes is the task of selecting one voice from of a set of voices. In forensic speaker identification the task and setting are similar (the set of voices might not include more than one voice), but the method does not have to operate in real-time (Rose 2003). That is, forensic speaker identification has fewer time constraints than more time critical automatic systems; the forensic speaker identification expert often has several days, or even weeks, to work with a problem, whereas a security system is expected to respond within seconds.

A common, yet not, universal assumption in most automatic systems is that the voice under investigation may not be one of the voices in the reference set. The concept of the open reference set, which presumes that the voice collected, the incriminating material, might not be part of the reference set of collected voices, should always be the working hypothesis in a forensic setting as a person is presumed innocent until proven otherwise.

The methods used in forensic voice identification and verification cases can be divided into three groups: aural-visual, automatic, and semi-automatic. Included in the aural-visual group is the much debated voiceprint analysis approach. The automatic and semi-automated methods differ from one another in that in the automatic approach there is no manual processing of the recording, where as in the semi-automatic some kind of manual pre-processing before automatic feature extraction and classification is performed.

A semi-automatic voice identifier based on spectral moments has been proposed (Rodman, McAllister, Bitzer, Cepeda, and Abbitt 2002) and tested for both English and Swedish (Eriksson, Cepeda, Rodman, McAllister, Bitzer, and Arroway 2004). This paper presents a pilot study that examines how vulnerable this method is to voice imitation attack – in effect a robustness test. The method is evaluated with both linear discriminant analysis and Mahalanobis distances. The paper first presents an overview of how this semi-automatic voice identification system operates. Then prior to a description of the classification methods, the speech data used are presented. This presentation includes an overview of the perception research that these speech samples have been used in, and outlines the impact of these speech samples upon human listeners undertaking speaker identification tasks. Finally, the results from the semi-automatic system's classifications are presented and discussed.

2. Method

The approach to speaker identification and discrimination used in this study is one that is currently under development at North Carolina State University (NCSU), USA. This method is used to extract parameters from speech that can be used with classification algorithms to determine a speaker's identity.

2.1. The NCSU approach to spectral moment extraction

To classify a specific target voice among a reference set of voices the method proposed in (Rodman et al. 2002) uses so called 'isochunks' (Eriksson et al. 2004). An isochunk is defined as a segment of speech that the speaker produces with essentially the same pronunciation each time it is produced and therefore sounds the same. For a segment to be used, more than one instance of the segment must exist in the recording of each speaker in the set of voices. An isochunk can be of arbitrary length and may contain linguistic boundaries; boundaries may, however, not introduce any pauses within the isochunk. The point is that an isochunk has the same underlying representation for which it might differ between speakers' ways of articulating it. Finally, the critical factor in the selection of the sound sequence as an isochunk is that the sequence selected should

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

sound as similar as possible within a certain speaker and as non-similar as possible between speakers.

After the isochunks have been selected and extracted the following algorithm is used to extract the features that are then used for classification.

1. Compute the discrete Fourier transform on a window of size N
2. Discard the imaginary part
3. Shift over one sample and repeat steps 1 and 2 N times.
4. Take the average of the N transforms and scale by the cube root; this lowers the impact of the first formant.
5. Interpolate the resultant average with a cubic spline to produce a pitch synchronous continuous spectrum.
6. Integrate the pitch synchronous continuous spectrum from 0 to 4000 Hz; this produces the mass of the spectrum (Eq. 1).

$$mass = \int_{0}^{4000} S(f)df \tag{1}$$

7. Divide the spectrum by its mass; this produces a probability density function (Eq. 2) and makes the area under the spectral curve essentially one and all sub-areas then lie in the range 0 – 1.

$$P(f) = \frac{S(f)}{mass} \tag{2}$$

8. Integrate the probability density function together with the frequency; this will yield the first moment, that is the mean of the function (see Eq. 3).

$$m_1 = \bar{x} = \int_{0}^{4000} f * P(f)df \tag{3}$$

9. Integrate the squared difference between the first moment and the frequency multiplied by the probability density function; this produces the second moment, that is the variation about the mean (Eq. 4).

$$m_2 = \sigma = \int_{0}^{4000} (f - m_1)^2 * P(f)df \tag{4}$$

10. Repeat steps 1 – 9 while the number of the samples left in the speech segment is more then $3N$.
11. Scale the first and second moment by 10^{-3} and 10^{-6} respectively.
12. Plot the first and second moment against each other. The first moment, the mean, is represented on the x-axis and the second moment, the variance about the mean, is represented on the y-axis. This forms a track.
13. Overlay the track with a minimal enclosing rectangle (MER). The final features for classification are extracted from the MER. Figs. 1 – 3 show examples of typical tracks with MERs overlayed.
14. Extract the following attributes from the MER: the minimum and maximum x and y values (each of these represents one corner), the lengths of the long and short sides, the x and y coordinates of the midpoint of the rectangle and the angle of orientation of the rectangle in relation to the x-axis.

2.2. Data

The speech material consisted of 10 male voices and was taken from the set of speech recordings used by (Zetterholm et al. 2002) and the set of recordings used by (Schlichting and Sullivan 1997). All recordings were of the same text, an excerpt from a political speech; the text and translation of this passage can be found in (Schlichting and Sullivan 1997) and (Sullivan and Schlichting 2000).

The data set includes the natural voice of a famous Swedish politician (ps), two professional imitations of this voice (amimit and ggimit) and three non-professional imitations of this voice (f1imit, f3imit and f5imit). The natural voices for the non-professional imitators and one of the professional imitators were also included as part of the data set (f1orig, f3orig, f5orig, amorig).

These particular voices were chosen based on their ability, or inability, to confuse human listeners in a set of perception experiments looking at the impact of voice imitation for speaker identification in forensic settings. Humans have been shown to exhibit less than perfect accuracy in detecting the 'right' voice in voice line-up experiments, even when the voice is well-known to the listener (Rose and Duncan 1995) and the sequence of perception experiments using the voices used in this study has shown that this lack of accuracy makes humans susceptible to voice imitation especially if the target voice is known (Zetterholm, Sullivan, Green, Eriksson, and Czigler 2003) and the semantic context is correct (Zetterholm, Sullivan, and van Doorn 2002) and (Sullivan et al. 2002).

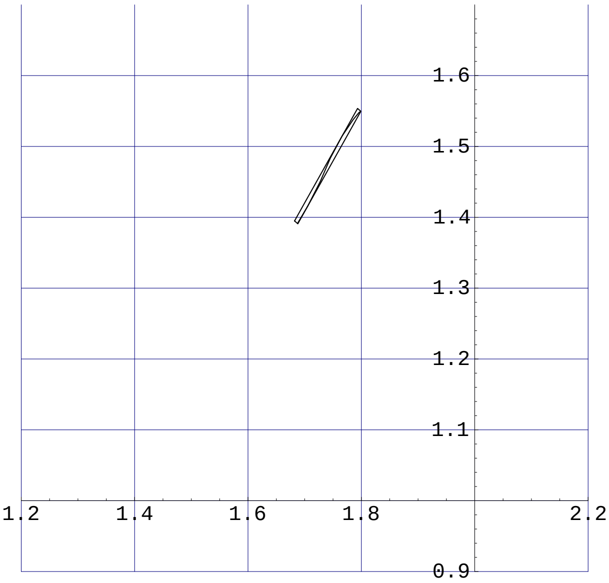

Figure 1: *Plot of the track formed by plotting the first and second moment against each other. Voice of the famous Swedish politician (PS) uttering /ljæ:/ in /miljæ:paʈiet/.*

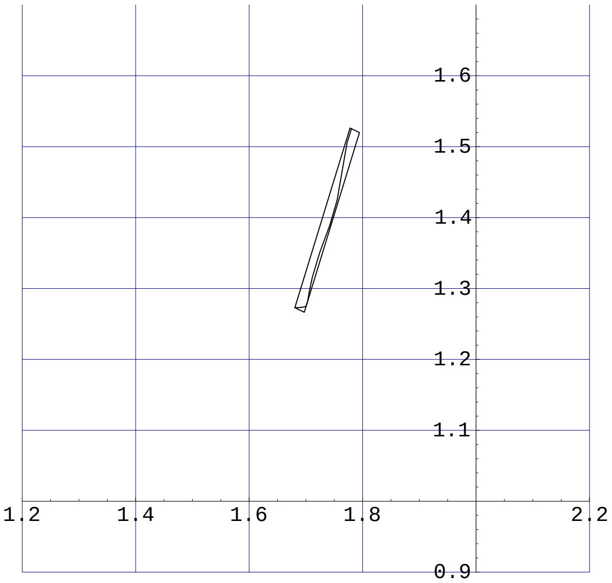

Figure 2: *Plot of the track formed by plotting the first and second moment against each other. Voice of one of the professional imitators during his imitation of the famous politician (amimit) uttering /ljæ:/ in /miljæ:paʈiet/.*

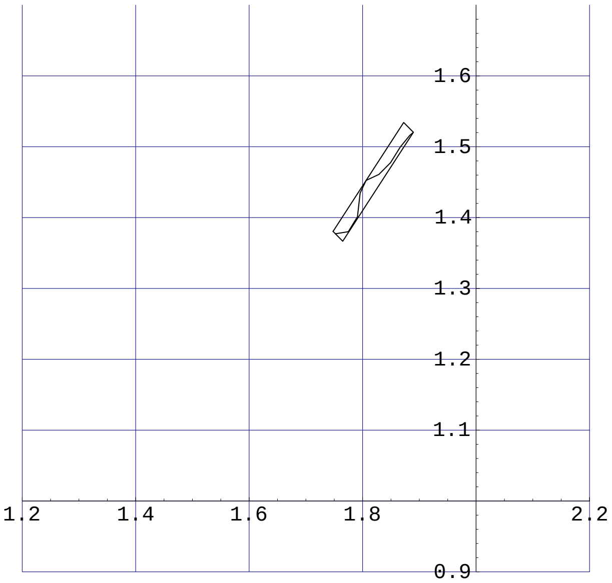

Figure 3: *Plot of the track formed by plotting the first and second moment against each other. The natural voice (amorig) of the imitator in Fig. 2 uttering /ljæ:/ in /miljæ:paʈiet/.*

The isochunk segment chosen to test the robustness of the spectral moment approach with imitation data was /ljæ:/, which qualifies as an isochunk by there being more than one good instance in each speaker's recording. Moreover, the length of this isochunk has been demonstrated to be long enough for the spectral moment approach (see (Eriksson et al. 2004) for a discussion on the length of an isochunk). The isochunks were carefully extracted prior to the extraction of their spectral moments and MERs.

Table 1: *Collapsed confusion matrix for the speakers assigned to the ten speaker models; where a speaker has more than one speaker model (i.e. a natural and an imitation model) these have been scored as if they were the same speaker model. Five isochunks were assigned one of ten speaker models (collapsed to six) for each of the ten voices. The voices are: two professional imitations (amimit and ggimit) the first professional imitator's natural voice (amorig), three amateur imitations (f1imit, f3imit, f5imit) and their natural voice counterpart (f1orig, f3orig and f5orig) and the target voice of the imitations, the politician's natural voice (PS).*

Voice	Collapsed speaker model					
	am	PS	gg	f1	f3	f5
amimit	1	1	0	0	0	3
amorig	1	3	0	0	1	0
PS	3	0	0	1	0	1
ggimit	0	1	4	0	0	0
f1imit	0	0	0	5	0	0
f1orig	2	0	0	3	0	0
f3imit	1	0	0	0	4	0
f3orig	0	0	0	0	5	0
f5imit	3	0	0	0	0	2
f5orig	0	0	0	0	0	5

2.3. Classification

Two approaches to classification have been applied. Linear discriminant analysis and Mahalanobis distances, both of which have previously been shown to be successful (Rodman et al. 2002).

Linear discriminant analysis creates a linear function using the parameters it is presented with and assumes a linear separation between the groups in the material. In this context, this is between the speakers. Each speaker is represented by a unique linear function. This function is created by calculating constants to assign to each parameter. Usually, about 90%, called the training data, of the data is used to construct the functions and the remaining 10% used to test. In a forensic setting, the available reference recordings would form the training data and the incriminating sample would form the test material. Here, however, the test data is the same as the training data; this was so that the validity for the selected isochunk and the extracted parameters could be validated.

Each segment in the test material is tested on every function and assigned membership to the function that yields the best result. The test material is then assigned to the function that has been assigned the greatest number of the segments in the material. This also presents a probability measure. It can be said that, for example, 10, 50, 80 or 100% of the cases in the test material were assigned to the function representing speaker X. The amount of material, both reference and testing, is critical to the statistical value of the probability measure. That is, if there are only a few cases of the segment in the test material, the classification will be very susceptible to errors, whereas if there is a large number of cases the classification will be less vulnerable.

Mahalanobis distances are calculated using the squared

Table 2: *Confusion matrix of the speaker assigned to the ten speaker models. The voices are: two professional imitations (amimit and ggimit) the first professional imitator's natural voice (amorig), three amateur imitations (f1imit, f3imit, f5imit) and their natural voice counterpart (f1orig, f3orig and f5orig) and the target voice of the imitations, the politician's natural voice (PS).*

Voice	Assigned speaker model									
	amimit	amorig	PS	ggimit	f1imit	f1orig	f3imit	f3orig	f5imit	f5orig
amimit	0	1	1	0	0	0	0	0	2	1
amorig	0	1	3	0	0	0	1	0	0	0
PS	0	3	0	0	0	1	0	0	0	1
ggimit	0	0	1	4	0	0	0	0	0	0
f1imit	0	0	0	0	2	3	0	0	0	0
f1orig	0	2	0	0	3	0	0	0	0	0
f3imit	0	1	0	0	0	0	0	4	0	0
f3orig	0	0	0	0	0	0	4	1	0	0
f5imit	3	0	0	0	0	0	0	0	0	2
f5orig	0	0	0	0	0	0	0	0	4	1

distance between instances in a multi-dimensional space and are defined for distance between each case and the midpoint of all cases. If the midpoint is changed to be any of the other cases, the distance between the two instances can be computed. By calculating multi-dimensional means of cases for each individual, the distance between two individuals can be acquired and therefore also the degree of separation between the individuals.

3. Results

The results for the linear discriminant analysis and Mahalanobis distances are presented separately.

3.1. Linear discriminant analysis

The linear discriminant analysis' result is presented in Tables 1 and 2 in the form of confusion matrices. It is important to remember that the test data is the same as the training data in order to check the validity of the selected isochunk and the parameters extracted. Ideal isochunk and parameters yield 100% correct classification rates. This would result in a value of 5 in Tables 1 and 2; there are five instances of the isochunk /ljœ:/ for each speaker.

3.2. Mahalanobis distances

The distances between each speaker are presented in Table 3. A T-test was performed between the distances of each speaker; the distances differing at a signification level are marked with a $*$ in Table 3.

4. Discussion

The results from the linear discriminant analysis presented in Tables 1 and 2, show a less than 100% successful identification rate. This is most likely due to the small size of the dataset; a larger set would, based on previous experience, have resulted in a better identification rate. Table 1 shows the collapsed confusion matrix of assigned speakers (in this table assignments are shown for the person). That is, the person's own voice and the person's imitation of the Swedish politician are considered to represent the same assigned individual. Examination of Table 1 shows that 100% is recorded for f1imit, f3orig and f5orig, and that the greatest degree of confusion is achieved by PS with a 0% correct identification rate, closely followed by the am voices. There is a clear confusion between the politician's voice and am's voices; this indicates a similarity between the voices, at least in regard to the MER values based on the isochunk /ljœ:/. The similarity detected here between PS and amorig, i.e. his natural voice and not his imitation of PS, could be a factor in this am's success in perception studies when imitating this Swedish politician's voice; the confusion found here between PS and amimit is interestingly lower.

The non-collapsed Table 2 provides a more detailed presentation of the assigned voices. This table shows how the majority of the incorrect assignments are within speaker, i.e., natural voice and imitation. The major exceptions are the confusion between PS and amorig, amimit and f5imit. Whether these confusions are due to perceptual similarities captured by the spectral moments of the isochunk or due to another factor demands further investigation with a larger database of isochunks.

The t-tests that were performed on the Mahalanobis distances (shown in Table 3) only found significant differences that involved the voice ggimit: no significant differences were found between a pair of voices that did not include ggimit as one of its voices. The results from the Mahalanobis distance measurements concur with those from the linear discriminant analysis. For example, the professional imitation, ggimit, is well discriminated and uniformly classified. Furthermore, the confusions found between speakers' natural voices and their imitations are revealed by the smaller Mahalanobis distances between these voices than between other pairs of voices. This can be see in the distance rankings shown for each voice in Table 3's columns; the voice that has the smallest Mahalanobis distance from the voice has rank 1 and the majority of these rankings are between the natural voice and the imitation by the same speaker, whereas, for example, ggimit is ranked 7 when compared with PS. For this it is possible to infer that the imitation was "seen through" by the system.

Table 3: *Absolute distances of mean Mahalanobis distances between each voice (D). Significant values at the* $p < .05$*-level are marked with an* $*$*. For each voice, the column R, shows the rank-order of the distances. The voices are: two professional imitations (amimit and ggimit) the first professional imitator's natural voice (amorig), three amateur imitations (f1imit, f3imit, f5imit) and their natural voice counterpart (f1orig, f3orig and f5orig) and the target voice of the imitations, the politician's natural voice (PS). The table is broken down into two parts where Table 4(a) is continued in Table 4(b).*

	amimit		amorig		PS		ggimit		f1imit	
Voice	D	R	D	R	D	R	D	R	D	R
amimit	–		89710.4	1	132038.0	6	269859.6	3	44632.3	1
amorig	89710.4	2	–		42327.6	4	359569.9*	5	45078.1	2
PS	132038.0	4	42327.6	3	–		401897.6*	7	87405.7	5
ggimit	269859.6	9	359569.9*	9	401897.6*	9	–		314491.8*	9
f1imit	44632.3	1	45078.1	4	87405.7	5	314491.8*	4	–	
f1orig	120250.0	3	30539.7	2	11788.0	1	390109.6*	6	75617.8	3
f3imit	236080.0	8	325790.3	8	368117.9	8	33779.6	1	280712.2	8
f3orig	212308.7	7	302019.0	7	344346.7	7	57550.9	2	256940.9	7
f5imit	149494.4	5	59784.1	5	17456.4	2	419354.0	8	104862.2	5
f5orig	169033.4	6	79323.1	6	36995.5	3	438893.0*	9	124401.2	6

(a)

	f1orig		f3imit		f3orig		f5imit		f5orig	
Voice	D	R	D	R	D	R	D	R	D	R
amimit	120250.0	6	236080.0	4	212308.7	3	149494.4	6	169033.4	6
amorig	30539.7	3	325790.3	6	302019.0	5	59784.1	4	79323.1	4
PS	11788.0	1	368117.9	8	344346.7	7	17456.4	1	36995.5	2
ggimit	390109.6*	9	33779.6	3	57550.9	2	419354.0	9	438893.0*	9
f1imit	75617.8	5	280712.2	5	256940.9	4	104862.2	5	124401.2	5
f1orig	–		356330.0	7	332558.7	6	29244.4	3	48783.4	3
f3imit	356330.0	8	–		23771.3	1	385574.4	8	405113.4	8
f3orig	332558.7	7	23771.3	2	–		361803.1	7	381342.1	7
f5imit	29244.4	2	385574.4	9	361803.1	8	–		19539.0	1
f5orig	48783.4	4	19539.0	1	381342.1	9	19539.0	2	–	

(b)

5. Conclusion

This pilot study suggests that the spectral moment approach to speaker identification and verification is not overtly sensitive to voice imitation; the 'incorrect' voice assignments by the spectral moment approach were not universally increased due to imitation. This could be, as pointed out above, due to the limited size of the dataset in this pilot study and the choice of isochunk. However, weaknesses with the approach have been found previously (Eriksson et al. 2004) and these cannot be ruled out from having an impact on the result. Ways to overcome these weaknesses are currently being investigated; improvements in the glottal pulse tracker and the possibility of using the curvature of the track in addition to the MER values to improve discrimination power are the focus of investigation at the moment. It it anticipated that spectral moments will become less sensitive to imitation attack after these improvements have been completed.

6. Acknowledgements

The Swedish part of this research was funded by the The Bank for Sweden Tercentenary Foundation through their funding of the project "Imitated voices: A research project with applications for security and the law." Dnr K2002-1121:1, Dnr K2002-1121:2. Gratitude is extended to North Carolina State University (NCSU) and the Department of Computer Science, NCSU, for their cooperation in the undertaking in this research.

References

Eriksson, E., L. Cepeda, R. D. Rodman, D. McAllister, D. Bitzer, and P. Arroway (2004). Cross-language speaker recognition using spectral moments. In *Proceedings FONETIK 2004, the XVIIth Swedish Phonetic Conference*, Stockholm, Sweden, pp. 76 – 79.

Labov, W. (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Rodman, R. D., D. McAllister, D. Bitzer, L. Cepeda, and P. Abbitt (2002). Forensic speaker identification based on specral moments. *Forensic Linguistics 9*(1), 22 – 43.

Rose, P. (2003). *Forensic Speaker Identification*. New York: Taylor & Francis.

Rose, P. and S. Duncan (1995). Naïve auditory identification and discrimination of similar voices by familiar listeners. *Forensic Linguistics: The International Journal of Speech, Language and the Law 2*, 1 – 17.

Schlichting, F. and K. P. H. Sullivan (1997). The Imitated Voice – a Problem for Voice Line-ups? *Forensic Linguistics: The International Journal of Speech, Language and the Law 4*, 129 – 132 (Errata 4: 318 – 319).

Sullivan, K. P. H. and F. Schlichting (2000). Speaker Discrimination in a Foreign Language: First Language Environment, Second Language Learners. *Forensic Linguistics: The International Journal of Speech, Language and the Law 7*, 1350 – 1771.

Sullivan, K. P. H., E. Zetterholm, J. van Doorn, J. Green, F. Kügler, and E. Eriksson (2002). The effect of removing semantic information upon the impact of voice imitation. In *The 9th Australian International Conference on Speech Science & Technology*, Melbourne, Australia.

Zetterholm, E., K. P. H. Sullivan, J. Green, E. Eriksson, and P. Czigler (2003). Imitation, expectation and acceptance: the role of age and first language in a nordic setting. In *ICPhS, The 15th International Congress of Phonetic Sciences*, Barcelona, Spain.

Zetterholm, E., K. P. H. Sullivan, and J. van Doorn (2002). The impact of semantic expectation on the acceptance of a voice imitation. In *The 9th Australian International Conference on Speech Science & Technology*, Melbourne, Australia.