# A TWO STAGE FUZZY DECISION CLASSIFIER FOR SPEAKER IDENTIFICATION

P. Castellano and S. Sridharan

Signal Processing Research Centre
School of Electrical and Electronic Systems Engineering
Queensland University of Technology

ABSTRACT - This paper discusses a new neural network based Two Stage Fuzzy Decision Classifier (TSFDC), applicable to Automatic Speaker Identification (ASI). At the conclusion of its training phase, this system computes one reference possibility set per speaker present in the data. During classification, these sets may be used to assist the network's decision making process in identifying a speaker for each input vector. The speaker is selected from amongst the two most likely speakers, for a given vector. The method was not found to be beneficial when identification scores already exceeded 54%. However, it consistently improved classification (by 15% on average) for those speakers harder to identify (below 54%). Misclassification by as much as 54% is uncommon for an ANN trained with a multiple speaker database. However, more work is needed to determine the probability of such an event occurring.

## INTRODUCTION

ASI uses the spectral characteristics of a person's voice to determine whether that person belongs or not to a reference database. Vector Quantisation and supervised Artificial Neural Networks (ANNs) have been widely applied to this pattern matching problem. While good results have been reported, these classifiers may perform poorly for some speakers. This is unfortunate since these speakers are often those most in need of identification.

This paper endeavours to improve the ASI performance of ANN classifiers when dealing with difficult speakers, in text independent mode. This work proposes a modification to these classifiers, by recognising that the single distance metric used to label vector patterns in terms of speaker classes is not a perfect classification criteria. The modified network incorporates a priori information into its decision making process.

## DATA

The present study considers two speaker databases composed of free conversational speech. They both contain low energy speech frames which are poor predictors of speaker identity (Rudasi and Zahorian, 1991). Database A (14 speakers) was compiled from recordings made over the radio. Database B (17 speakers each) was derived from audio-visual material. Both databases were composed of voices from middle aged Anglo-Saxon speakers.

The speech signal was digitised at 10 kHZ (12 bits). It consisted of 50 second segments with silent parts removed. The segments were divided into frames which were first high-frequency pre-emphasised with transfer function $1 - 0.98z^{-1}$. Other pre-processing specifications include the use of a non overlapping 256 point Hamming window and an analysis filter of order 15. Reflections coefficients were computed for database A and mel-based cepstra for B. For each database and speaker, two distinct sets of 100 vectors were set aside for the purpose of respectively training and testing classifiers.

## ANN CLASSIFIERS FOR SPEAKER IDENTIFICATION

There are two stages in the operation of ANNs namely training and classification. During training, the connection weights are modified in response to input data vectors. If the learning is supervised, target vectors are simultaneously presented at the output, expressing a desired response to the given input. During classification, a network processes information and outputs a response. Output is classified

according to which target it most closely approximates. The present study is based on a single ANN with class specific output.

## TWO STAGE FUZZY DECISION CLASSIFIER (TSFDC)

Whether in training mode or classification mode, the TSFDC's ANN produces one output response for each input vector presented to it (recall). Within the ASI context, the ith response, for speaker j, is expressed as

$X_{j,i}:\{[y_1, y_2, ..., y_k, .. y_n]\}$

where n is the number of speakers studied, $(0< j < n+1)$, $(0< k < n+1)$ and $y_k$ is on [0,1).
A TSFDC depends on each $y_k$ conforming to the concept of fuzzy set discussed elsewhere (Zadeh, 1987). Let $\{s_k : S\}$ where S is the set of all n speakers considered. For output i and speaker j, each $y_k$ is expressed as

$$y_k = \text{Poss}\,(X_{j,i}, s_k) \qquad (1)$$

where Poss $(X_{j,i}, s_k)$ is the possibility that $X_{j,i}$ may be identified with speaker $s_k$. In this context, possibility refers to the degree of ease with which the identification is made and

$$\sum_{k=1}^{n} y_k \neq 1 \qquad (2)$$

for a given response and speaker. In agreement with Zadeh's concept, $X_{j,i}$ is a possibility set expressing the degree by which $X_{j,i}$ can be identified with each one of the n speakers. The set of $y_k$s, derived over all output responses, for a given speaker and k, is a possibility distribution. Therefore the set of all ANN training responses, for a given speaker, is an n dimensional possibility distribution (n-DPD). For easily classified speech vectors, equ.(2) is almost an equality and hence almost probabilistic. For those vectors, $y_k$s are all close to zero except for one $y_k$ which is close to 1. The position of the latter $y_k$ identifies the speaker. For other speech vectors, several $y_k$s may approach 1 making classification more difficult.

### Creating reference possibility sets

TSFDC produces one n-DPD, for each speaker, from the last iteration through the training data. If training has been successful, these distributions are associated with the smallest global network error. TSFDC automatically quantises the n-DPDs to produce one reference (possibility) set per speaker. This quantisation is done by taking the arithmetic mean of each possibility distribution in an n-DPD.

### TSFD classification

The reference possibility sets, produced at the end of training, represent a priori information. This information can be used in conjunction with target information to classify, in two stages (TSFD), the output produced by the ANN during an application. At the beginning of the classification phase, the user is prompted for TSFD. If TSFD is not selected, the system's mode of operation is identical to that of a lone ANN. The winning speaker class, for each output vector, is identified with the position of the greatest value in that vector. This is a one stage fuzzy classification. With TSFD selected, the greatest and second greatest values (valG and valSG) are found. For an n speaker problem and a given ANN output vector, these values are given by

$$\text{valG} = \max(y_k) \qquad (3)$$
$$\text{valSG} = \text{second } \max(y_k) \qquad (4)$$

where k $\in$ {1,..,n}. Their respective positions (positionG and positionSG) and associated reference set values (RefG and RefSG) are retained. During the second stage of classification, the value at positionG in RefSG ($valG_{RefSG}$) and that at positionSG in RefG ($valSG_{RefG}$) are inspected. Finally, two weighted Euclidean distance metrics are computed. These are given by

$$SUM1 = | upper\_b - valG | + W * | valSG_{RefG} - lower\_b | \quad (5)$$
$$SUM2 = | upper\_b - valSG | + W * | valG_{RefSG} - lower\_b | \quad (6)$$

where upper_b and lower_b are respectively the upper (1) and lower (0) bounds of the value space. W is a weight set heuristically and is also real. W governs the degree by which the vector quantisation of the training outputs is allowed to influence the TSFDC's final classification. If SUM2 is greater or equal to SUM1 the speaker class corresponding to valG is declared the winner. The winner is valSG otherwise. Classification may be executed several times both with and without TSFD. Since the ANN need not be re-trained between classifications, the overhead is negligible. This enables the TSFDC to be easily bench marked against the performance of its own ANN for various Ws. The TSFDC's mode of operation is summarised in Figure 1.

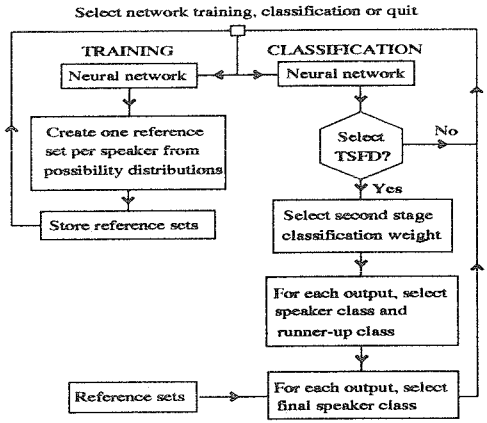


Figure 1. Two Stage Fuzzy Decision Classifier.

This study investigates 2 ANN architectures which can be included into the TSFDC. The first is the Multi-layer Perceptron (MLP) widely used in ASI (Rudasi and Zahorian, 1991). With TSFD, the number of output nodes and the dimension of the target must equal the number of speakers present in the speech signal. Target vector values are chosen to be 0.999 for the correct class and 0.001 for all others. The second architecture considered is the Tensor based Functional-link Network also known as Higher Order Neural Network ( HONN) (Hush and Salas, 1989). With HONN, each input vector $[x_1, x_2, ..., x_n]$ is first mapped into a higher dimensional vector $[x_1, x_2, x_3, ..., x_n, x_1x_2, x_1x_3, x_1x_2x_3, ..., x_{n-2}x_{n-1}, x_{n-2}x_n, x_{n-2}x_{n-1}x_n]$. Target values are the same as for the MLP.

## RESULTS AND DISCUSSION

### TSFDC with MLP (Database A: speakers A to N)

The MLP had 15 inputs, one layer of 40 hidden units and 14 output units. Each output vector corresponding to a given speaker (say speaker J) was inspected to determine whether the value in the position for that speaker ($y_j$) was either the greatest (correct classification) or second greatest (near miss). Results are shown in Figure 2.

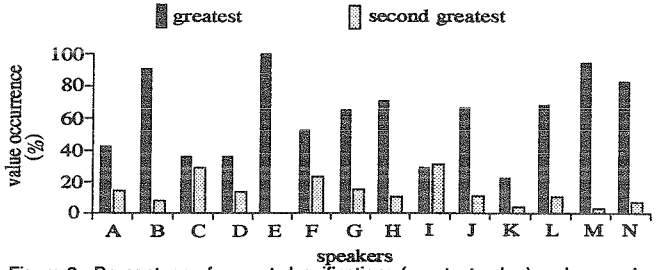


Figure 2. Percentage of correct classifications (greatest value) and near misses (second greatest) in MLP output for speakers in Database A.

Near misses did not degrade the MLP's performance uniformly across speakers. Near misses were rare for those speakers which the MLP could unambiguously recognise on its own (speakers B, E and M). Tables 1 to 3 show results for the TSFDC given different values of W (speakers A to N).

Table 1. Confusion matrix for Database A (W=0).

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 44 | | | 10 | 14 | | | 9 | | 6 | | 1 | 9 | 7 |
| B | | 91 | | 1 | | | | 2 | | 2 | | 1 | | 3 |
| C | | 2 | 38 | 2 | 5 | 14 | 3 | 26 | 3 | 1 | 3 | 3 | | |
| D | 10 | 1 | | 39 | 7 | 2 | | 20 | | 4 | 2 | 2 | 5 | 8 |
| E | | | | | 100 | | | | | | | | | |
| F | | 1 | 5 | | 6 | 55 | 8 | 17 | 1 | | 2 | 4 | 1 | |
| G | | | | | 9 | 9 | 65 | | 10 | | 2 | | 5 | |
| H | | | 12 | 1 | 8 | 2 | | 72 | | 1 | 2 | 1 | | 1 |
| I | | | 1 | | 7 | 25 | 26 | 1 | 29 | 1 | 4 | 2 | 4 | |
| J | 4 | | | 2 | 3 | | | 6 | | 68 | 2 | | 9 | 6 |
| K | 6 | 2 | 10 | 9 | 3 | 4 | 5 | 26 | 2 | 5 | 23 | | 3 | 2 |
| L | 1 | 3 | | | 2 | 14 | | 5 | 3 | | | 72 | | |
| M | 1 | | | 1 | | | | | | | | | 97 | 1 |
| N | 4 | | | 3 | | | | 3 | | 2 | | | 6 | 82 |

Table 2. Confusion matrix for Database A (W=10).

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 47 | | | 10 | 16 | 1 | | 7 | | 7 | | 1 | 6 | 5 |
| B | | 91 | | 2 | | | | 2 | | 2 | | 1 | | 2 |
| C | | 3 | 42 | 5 | 6 | 14 | 2 | 20 | 1 | 2 | 3 | 2 | | |
| D | 10 | 2 | | 41 | 8 | 2 | | 20 | | 4 | 1 | 1 | 4 | 7 |
| E | | | | | 100 | | | | | | | | | |
| F | | 1 | 5 | | 5 | 54 | 7 | 17 | 5 | | 1 | 4 | 1 | |
| G | | | | | 10 | 8 | 60 | | 14 | 1 | 2 | | 5 | |
| H | | 1 | 13 | 1 | 6 | 3 | | 72 | | 1 | 1 | 1 | | 1 |
| I | | | | | 6 | 21 | 17 | 1 | 43 | 1 | 4 | 2 | 5 | |
| J | 3 | | | 1 | 6 | | | 5 | | 70 | 2 | | 6 | 7 |
| K | 7 | 3 | 9 | 11 | 3 | 4 | 3 | 23 | 3 | 6 | 24 | | 2 | 2 |
| L | 1 | 3 | | 1 | 1 | 13 | | 7 | 3 | | | 71 | | |
| M | 1 | | | 1 | | | | | | | 1 | | 96 | 1 |
| N | 6 | | | 5 | | | | 3 | | 2 | | | 4 | 80 |

Table 3. Confusion matrix for Database A (W=30).

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 46 | | | 11 | 15 | 1 | | 7 | | 7 | | 2 | 6 | 5 |
| B | | 89 | | 2 | | | | 2 | | 2 | | 3 | | 2 |
| C | | 4 | 42 | 5 | 6 | 14 | 2 | 21 | 1 | 2 | 1 | 2 | | |
| D | 10 | 2 | | 40 | 8 | 4 | | 18 | | 4 | 1 | 1 | 4 | 8 |
| E | | | | | 100 | | | | | | | | | |
| F | | 1 | 4 | | 4 | 56 | 6 | 16 | 6 | 1 | 1 | 4 | 1 | |
| G | | | | | 11 | 5 | 41 | | 35 | 1 | 2 | | 5 | |
| H | | 1 | 15 | 4 | 6 | 3 | | 68 | | 1 | 1 | | | 1 |
| I | | | | | 6 | 18 | 11 | 1 | 51 | 1 | 4 | 2 | 6 | |
| J | 3 | | | 1 | 6 | | | 5 | | 71 | 1 | | 5 | 8 |
| K | 7 | 3 | 10 | 11 | 3 | 5 | 2 | 23 | 2 | 6 | 24 | | 2 | 2 |
| L | 1 | 3 | | 3 | 1 | 10 | | 7 | 6 | | | 69 | | |
| M | 1 | | | 1 | | | | | | | 1 | | 95 | 2 |
| N | 7 | | | 8 | | | | 3 | | 3 | | | 3 | 76 |

As W was increased, ASI improved for speakers with initially low identification scores ( below 54 percent for W=0). On a speaker by speaker basis, improvement in classification appeared to be influenced by the proportion of near misses to correct ASI shown in Figure 2. For instance speaker I recorded a 22 point jump in ASI accuracy, as W was increased from 0 to 30 while speaker K went up by 1 through the same weight range. ASI scores did not improve for the initially better speakers, as W was made greater. The mean ASI score increased by a non significant 1 percent with a change in W from 0 to 10. At the same time the correct identification scores initially below 54% increased by an average of 14%. No speaker was mistaken for another for as many as 54% of speech vectors.

Figure 3 shows the 14 reference sets obtained from quantising all n-DPDs generated by the last iteration through the training data. Intra speaker peaks swamp in magnitude other vector components. W must be large for the latter components to significantly influence SUM1 equ.(5) and SUM2 equ.(6).

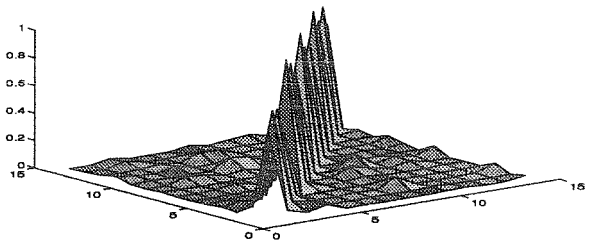

Figure 3. Reference possibility sets for 14 speakers (Databse A).

TSFDC with HONN (Database B: speakers A to Q)

The HONN had 15 inputs expanded into 15x4 higher order terms. The output layer contained 17 units. A high proportion of near misses could be identified for some speakers. See Figure 4.

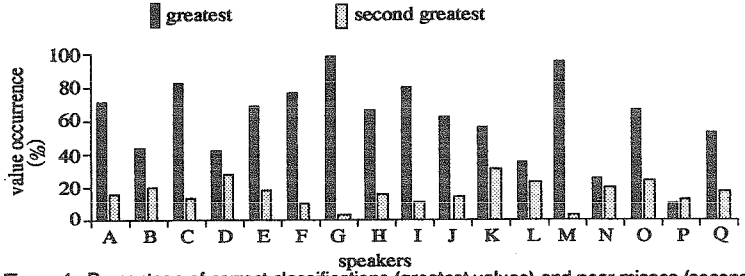


Figure 4. Percentage of correct classifications (greatest values) and near misses (second greatest) in HONN output for speakers in Database B.

This situation resulted in an ASI threshold ( lowest identification score in a speaker set) of only 9 when HONN was used alone (Table 4). Varying W from 0 to 10 (Tables 4 and 5) almost doubled that threshold but still resulted in misclassification for speaker N. All ASI scores, initially at and below 56%, were improved. This was accompanied by an increase in mean ASI score of 3%. Classification improved by an average of 16% for speakers with correct identification scores initially below 54%.

Table 4. Confusion matrix for Database B (W=0).

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 70 | | 5 | 1 | 1 | | 7 | | 8 | 3 | | | 1 | | 1 | | 3 |
| B | | 46 | 1 | | 1 | 18 | | 1 | | 5 | | 1 | | | 27 | | |
| C | 4 | | 82 | | | | | | 5 | 1 | | | 8 | | | | |
| D | | | | 44 | | 40 | 7 | | 4 | 2 | 1 | | | | | | 2 |
| E | | | | | 67 | | 1 | | 15 | 5 | 4 | | | | 4 | | 2 |
| F | | 1 | | | | 75 | 2 | | 3 | 4 | | | 10 | | | | 5 |
| G | | | | | | 1 | 96 | | | 1 | | | | | 2 | | |
| H | | 12 | 1 | 1 | | 3 | | 67 | | 7 | | | | | 9 | | |
| I | | | | 3 | | 5 | | | 80 | 3 | 5 | | | | | | 4 |
| J | 1 | | | 1 | 5 | 8 | 8 | 1 | 2 | 63 | 2 | | | | 2 | | 7 |
| K | | | 6 | 1 | 1 | | | | 36 | | 55 | | | | 1 | | |
| L | | 6 | | | 9 | 6 | 14 | | 1 | 2 | | 36 | 1 | 2 | 23 | | |
| M | | | 2 | | | 1 | 1 | | 2 | | | | 92 | | 2 | | |
| N | 2 | | 27 | | 1 | | 6 | | 4 | 14 | | 1 | 7 | 27 | 11 | | |
| O | | 1 | | | 4 | 16 | 1 | | | 7 | | | 1 | | 70 | | |
| P | 1 | | 5 | | 8 | 2 | 3 | | 30 | 7 | 23 | 1 | 1 | | 2 | 9 | 8 |
| Q | 2 | | | 5 | 6 | | 11 | | 15 | 2 | 3 | | | | | | 56 |

Table 5. Confusion matrix for Database B (W=7&10).

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 78 | | 3 | 1 | | | 6 | | 4 | 2 | | | 1 | | 1 | | 4 |
| B | | 52 | | | 1 | 20 | | 2 | | 4 | | | | | 21 | | |
| C | 4 | | 72 | | | | | | 11 | 1 | 1 | | 11 | | | | |
| D | | | | 48 | | 34 | 7 | | 4 | 2 | 1 | | | | | | 4 |
| E | 2 | | | | 45 | | | | 27 | 8 | 6 | | | | 5 | | 7 |
| F | | | | | | 74 | 1 | | 3 | 3 | | | 13 | | | | 6 |
| G | | | | 1 | | 8 | 88 | | | 1 | | | | | | | 2 |
| H | | 12 | 1 | 1 | | 5 | | 73 | | 3 | | | | | 5 | | |
| I | 1 | | | 2 | | 5 | | | 72 | 2 | 11 | | | | | | 7 |
| J | 4 | | | | 3 | 8 | 7 | 3 | 3 | 56 | 3 | | | | 2 | | 11 |
| K | | | 1 | | 1 | | | | 21 | | 75 | | | | 1 | | 1 |
| L | | 12 | 1 | | 4 | 4 | 14 | | 2 | 3 | | 40 | 2 | 2 | 16 | | |
| M | | | | | | | 1 | | 2 | | | | 95 | | 2 | | |
| N | 4 | | 20 | | 3 | | 5 | | 8 | 9 | | 1 | 7 | 34 | 9 | | |
| O | | 5 | | | 4 | 20 | 1 | 2 | | 7 | | | 1 | | 60 | | |
| P | 1 | | 4 | | 5 | 2 | 3 | | 22 | 12 | 24 | 1 | 1 | | 1 | 17 | 7 |
| Q | 2 | | | 5 | 4 | | 11 | | 11 | 1 | 7 | | | | | | 59 |

### Composite ASI results

Because several classification phases may be run sequentially, with TSFDC each time an option, it is a straight forward matter to generate a first set of results from the ANN alone then a second set with the TSFD operating (with W = 10). Only the ASI scores of those speakers who can be unambiguously identified need be retained from the first set. The second set yields the remaining (improved) scores. The threshold, set to separate ambiguous from unambiguous speaker identifications, is a matter of adjudication. For the purpose of this study, this threshold has been set at 54% correct identifications.

## CONCLUSION

Supervised connectionist networks rely solely on a comparison between output and target to classify that output into a single class. However, in some applications, the output is fuzzy and is therefore identifiable, to varying degrees, with all classes. We have proposed a Two Stage Fuzzy Decision Classifier which consists of an ANN with fuzzy output and a new discrimination algorithm. Given an n class problem, an n dimensional possibility distribution is computed for each class, after the ANN's last training iteration. (A distribution corresponds to the ANN's response to the last presentation of the input set.) One reference set is then derived by quantisation from each such distribution. A first stage of classification identifies the two most likely classes. In a second stage, a winner is picked from amongst the two. The second classification stage is governed both by distance metrics computed in stage one and metrics resulting from a consideration of the reference sets.

The applicability of the Two Stage Fuzzy Decision Classifier to text-independent ASI problems was investigated in this research. The classifier was found to improve the scores of all speakers most prone to misclassification by an average of 15%. However, the scores of some of the better speakers were reduced. This compromise resulted irrespective of the type of acoustic parameter used (Reflection or Cepstral). It can be avoided by opting for two independent classifying phases. Speaker reference sets are not used during the first phase. Hence, in this case, the operation of the classifier is that of a standard ANN. During the second phase, the reference sets are taken into account. Only the best ASI scores obtained from each classification phase are retained. Because no retraining of the ANN is required the overhead resulting from a two phase classification is negligible. Misclassification by as much as 54% is uncommon for an ANN trained with a multiple speaker database. However, more work is needed to determine the probability of such an event occuring.

## ACKNOWLEDGMENTS

This research was supported by grants from the National Institute of Forensic Science, Australia and by a Research Encouragement Award from the Queensland University of Technology.

## REFERENCES

Hush D. R. and Salas J. M. (1989) "Classification with Neural Networks: A Comparison", ISE '89, Eleventh Annual Symposium - Economic Growth through Science and Engineering, Ideas in Science and Electronics, 107-114.

Rudasi, L. and Zahorian, S. A. (1991) "Text-Independent Talker Identification with Neural Networks", Proc. of ICASSP, Vol. 1, 389-392.

Zadeh, L. A. (1987) "Commonsense and Fuzzy Logic", in The Knowledge Frontier - Essays in the Representation of Knowledge, ed. by N. Cercone and G. McCalla (Springer-Verlag New York Inc.), Chap. 5, 338-353.