FUZZY-INTEGRATION BASED NORMALIZATION FOR SPEAKER VERIFICATION

Tuan Pham, Michael Wagner

University of Canberra
Faculty of Information Sciences and Engineering
ACT 2601, Australia
E-mail: tuanp@ise.canberra.edu.au

ABSTRACT

Similarity normalization techniques are important for speaker verification systems as they help to better cope with speaker variability. In the conventional normalization, the *a priori* probabilities of the cohort speakers are assumed to be equal. From this standpoint, we apply the theory of fuzzy measure and fuzzy integral to combine the likelihood values of the cohort speakers in which the assumption of equal *a priori* probabilities is relaxed. This approach replaces the conventional normalization term by the fuzzy integral which acts as a non-linear fusion of the similarity measures of an utterance assigned to cohort speakers. Experimental results show that the speaker verification system using the fuzzy integral is more flexible and favorable than the conventional method.

1. INTRODUCTION

In speaker verification systems, the normalization techniques are important as they help to alleviate the variations in the speech signals, which are due to noise, different recording and transmission conditions [1]. There are two types of normalization techniques for speaker recognition: *parameter* and *similarity*. Some typical works in the parameter type were proposed by Atal [2], Furui [3]; and in the similarity type were by Higgins *et al.* [4], Matsui and Furui [5]. It has also been reported that most of speaker verification systems are based on the similarity-domain normalization [6]. We therefore, in this paper, will focus our attention to the verification mode with respect to the similarity normalization.

Generally in most similarity normalization techniques, the likelihood values of the utterance from the cohort speakers whose models are closest to the claimant model, are assumed to be equally likely. In reality, however, this assumption is not often true as the similarity measures between each cohort speaker and the client speaker may be different. Basing our motivation on this drawback, we introduce a new normalized log-likelihood method using the concept of fuzzy fusion. We relax the assumption of equal likelihood by imposing the fuzzy measures of the similarities between the cohort speaker models and the client model. Then the scoring of the cohort models can be obtained by the fuzzy integral which acts as a fusion operator with respect to the fuzzy measures.

2. SIMILARITY-DOMAIN NORMALIZATION

Given an input set of speech feature vectors
$X = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_N\}$,
the verification system has to decide if X was spoken by the client (for the sake of simplicity, from now on we will denote $\vec{x}$ as x). Based on the similarity domain, this can be seen as a statistical test between $H_0 : S$ and $H_1 : S'$ where H_0 is the null hypothesis that the claimant is the client S, while H_1 is the alternative hypothesis that the claimant is an impostor S'. The decision according to the Bayesian rule for minimum risk is given by

$$L(X) = \frac{p(X|S)}{p(X|S')} \begin{cases} > \theta & : \quad X \in H_0 \\ \leq \theta & : \quad X \in H_1 \end{cases} \qquad (1)$$

where θ is a prescribed threshold. Taking the logarithm, the likelihood ratio of (1) becomes

$$\log p(X|S) - \log p(X|S') \begin{cases} > \log\theta & : \quad X \in H_0 \\ \leq \log\theta & : \quad X \in H_1 \end{cases} \qquad (2)$$

where $\log L(X)$ is also called the normalized log-likelihood score. The normalized log-likelihhood value of X given the client model can be determined as

$$\log p(X|S) = \frac{1}{N} \sum_{n=1}^{N} \log p(x_n|S). \qquad (3)$$

Two common methods called the *geometric mean* and the *maximum* [7] can be used to calculate the normalized log-likelihood score given not the client model. For a set of background speaker models of size B: $S' = \{S_1, S_2, \ldots, S_B\}$, the geometric mean method is defined as

$$\log p(X|S') = \frac{1}{B} \sum_{b=1}^{B} \log p(X|S_b). \qquad (4)$$

The maximum method is defined as:

$$\log p(X|S') = \max_{S_b \in S'} \log p(X|S_b\}. \tag{5}$$

where the term $\log p(X|S_b)$ in both (4) and (5) can be calculated as in (3), and except for the scale $1/N$, it is the probability of an utterance X coming from one of the cohort speakers with the assumption that the *a priori* probabilities being equal.

As the main purpose of this paper is to attempt to improve the scoring of the similarity normalization, we will simply use the vector quantization (VQ) method to generate the acoustical models. Thus, the log-likelihood in terms of the VQ distortion measure between the set of training vectors X of the claimed speaker and the codebook of a speaker S can be expressed as

$$\log p(x_n|S) = -\min_k [D(x_n, b_k(S)], \ k = 1, 2, \ldots, K \tag{6}$$

where $x_n \in X$, $b_k(S)$ is a codeword of speaker S, and K is the codebook size.

3. FUZZY MEASURE AND FUZZY INTEGRAL

Stemming from the concept of fuzzy sets proposed by Zadeh [12], Sugeno developed the notions of fuzzy measure and fuzzy integral [8]. A fuzzy measure is a set function with monotonicity but not always additivity, and a fuzzy integral is a functional with monotonicity which is used for aggregating information from multiple sources with respect to the fuzzy measure.

3.1. Fuzzy measure

Let Y be an arbitrary set, and $\mathcal{B}$ be a Borel field of Y. A set function g defined on $\mathcal{B}$ is a fuzzy measure if it satisfies the following three axioms:

1. Boundary conditions: $g(\emptyset) = 0, g(Y) = 1$.

2. Monotonicity: $g(A) \leq g(B)$ if $A \subset B$, and $A, B \in \mathcal{B}$.

3. Continuity: $\lim_{i\to\infty} g(A_i) = g(\lim_{i\to\infty} A_i)$ if $A_i \in \mathcal{B}$ and $\{A_i\}$ is monotone (an increasing sequence of measurable sets).

A g_λ-fuzzy measure is also proposed by Sugeno which satisfies another condition known as the λ-rule ($\lambda > -1$):

$$g(A \cup B) = g(A) + g(B) + \lambda g(A) g(B),$$

where $A, B \subset Y$, and $A \cap B = \emptyset$.

It is noted that when $\lambda = 0$, the g_λ-fuzzy measure becomes a probability measure. In general, the value of the constant λ can be determined by the properties of the g_λ-fuzzy measure as follows.

Let $Y = \{y_1, y_2, ..., x_m\}$. If the fuzzy density of the g_λ-fuzzy measure is defined as a function $g : y_i \in Y \to [0, 1]$ such that $g_i = g_\lambda(\{y_i\})$, $i = 1, 2, ..., m$, then the g_λ-fuzzy measure of a finite set can be obtained as [10]

$$g_\lambda(Y) = \sum_{i=1}^{n} g_i + \lambda \sum_{i_1=1}^{m-1} \sum_{i_2=i_1+1}^{n} g_{i_1} g_{i_2} + ... + \lambda^{m-1} g_1 g_2 ... g_m. \tag{7}$$

Provided that $\lambda \neq 0$, (7) can be rewritten as

$$g_\lambda(Y) = \frac{1}{\lambda} \left[\prod_{i=1}^{m} (1 + \lambda g_i) - 1 \right]. \tag{8}$$

With boundary condition $g(Y) = 1$, the constant λ can be determined by solving the following equation:

$$\lambda + 1 = \prod_{i=1}^{m} (1 + \lambda g_i). \tag{9}$$

3.2. Fuzzy Integral

Let $(Y, \mathcal{B}, g)$ be a fuzzy measure space and $f : Y \to [0, 1]$ be a $\mathcal{B}$-measurable function. A fuzzy integral over $A \subset Y$ of the function f with respect to a fuzzy measure g is defined by

$$\int_A f(y) \circ g(\cdot) = \sup_{\alpha \in [0,1]} [\min(\alpha, g(f_\alpha))] \tag{10}$$

where f_α is the α level set of f, $f_\alpha = \{y : f(y) \geq \alpha\}$.

The fuzzy integral in (10) is called the Sugeno integral. When $Y = \{y_1, y_2, ..., y_n\}$ is a finite set, and $0 \leq f(y_1) \leq f(y_2) ... \leq f(y_n) \leq 1$, (if not, the elements of Y are rearranged to make this relation hold), the Sugeno integral can be computed by

$$\int_A f(y) \circ g(\cdot) = \max_{i=1}^{m} [\min(f(y_i), g(A_i))] \tag{11}$$

where $A_i = \{y_i, y_{i+1}, ..., y_m\}$, and $g(A_i)$ can be recursively calculated in terms of the g_λ-fuzzy measure as

$$g(A_i) = g_i + g(A_{i-1}) + \lambda g_i g(A_{i-1}), \quad 1 < i \leq m. \tag{12}$$

It can be seen that the above definition is not a proper extension of the usual Lebesgue integral, which is not recovered when the measure is additive. In order to overcome this drawback, the so-called Choquet integral was proposed by Murofushi and Sugeno [10]. The Choquet integral of f with respect to a fuzzy measure g is defined as follows:

$$\int_A f(y) dg(\cdot) = \sum_{i=1}^{m} [f(y_i) - f(y_{i-1})] \, g(A_i) \tag{13}$$

in which $f(y_0) = 0$.

4. FUZZY-FUSION BASED NORMALIZATION

It has been mentioned in the foregoing sections that the *a priori* probability of an utterance given that it is from one of the cohort speakers is assumed to be equal in the conventional similarity normalization methods, we use the concept of the fuzzy measure to calculate the grades of similarity or closeness between each cohort speaker model and the cliend model, ie. the fuzzy density, and the multi-attributes of these fuzzy densities. The final score for the normalization of the cohort speakers can then be determined by combining all of these fuzzy measures with the corresponding likelihood values using the Choquet integral. We express the proposed model in mathematical terms as

$$\log L(X) = \log p(X|S) - \log F(X|S') \qquad (14)$$

where $F(X|S')$ is the fuzzy integral of the likelihood values of an utterance X coming from the cohort speaker set $S' = \{S_b : b = 1, 2, \ldots, B\}$ with respect to the fuzzy measures of speaker similarity. It is defined as follows:

$$F(X|S') = \sum_{b=1}^{B} [p(X|S_b) - p(X|S_{b-1})]\, g(Z_b) \qquad (15)$$

where $p(X|S_b)$ has been previously defined,
$Z_b = \{S_b, S_{b+1}, \ldots, S_B\}$, $g(Z_b)$ is the fuzzy measure of Z_b, $p(X|S_0)$ = 0, and the relation
$0 \leq p(X|S_1) \leq p(X|S_2), \ldots, p(X|S_B)$ holds, otherwise the elements in S' need to be rearranged.

From the previous presentation of the fuzzy measure and the fuzzy integral, it is noticed that the key factor for the fuzzy fusion process is the fuzzy density. If the fuzzy densities can be determined then the fuzzy measures can be identified, which make it ready for the operation of the fuzzy integral. For the fusion of similarity measures, we consider the fuzzy density as the degree of similarity or closeness between the acoustic model of a cohort speaker and that of the client, ie. the greater the value of the fuzzy density is, the closer the two models are. Therefore, we define the fuzzy density as

$$g_b = 1 - \exp(-\alpha ||\vec{v}_B - \vec{v}_S||^2) \qquad (16)$$

where α is a positive constant, $||.||^2$ is the Euclidean norm which indicates the root-mean-square averaging process, $\vec{v}_b$ is the mean code-vector of a cohort speaker S_b, and $\vec{v}_S$ is the mean code-vector of the client speaker S.

It is reasonable to assume that some acoustic models of a cohort speaker, say S_1, may be more similar to those of the client speaker S than those of another cohort speaker, say S_2. However, some other acoustic models of S_2 may be more similar to those of S than those of S_1. Since the mean code-vectors are globally generated from the codebooks including all different utterances of the speakers, we therefore introduce the constant α in (16) for each cohort speaker in order to fine-tune the fuzzy density with respect to the Euclidean distance measure. At present we select the values of α by means of the training data and will further discuss this issue in the experimental section.

5. EXPERIMENTS

The commercial TI46 speech data corpus is used here for the experiments. The TI46 corpus contains 46 utterances spoken repeatedly by 8 female and 8 male speakers, labeled f1-f8 and m1-m8, respectively. The vocabulary contains a set of 10 computer commands: {*enter, erase, go, help, no, rubout, repeat, stop, start, yes*}. Each speaker repeated the words 10 times in a single training session, and then again twice in each of 8 testing sessions. The corpus is sampled at 12500 samples/s and 12 bits/sample. The data were processed in 20.48 ms frames at a frame rate at 125 frames/s. The frames were Hamming windowed and preemphasized with μ=0.9. 46 mel-spectral bands of a width of 110 mel and 20 mel-frequency cepstral coefficients (MFCC) were determined for each frame.

In the training session, each speaker's 100 training tokens (10 utterances x 1 training session x 10 repetitions) were used to train the speaker-based VQ codebook by clustering the set of all the speakers' MFCC into codebooks of 32, 64 and 128 codewords using the LBG algorithm [12].

The verification was tested in the text-dependent mode. Since both the geometric mean and the fuzzy fusion methods operate on the principle of integration and depend on the size of the cohort set, we therefore compare the performances of these two methods. This is a closed set test as the cohort speakers in the trainig are the same as those in the testing. For the purpose of comparison and due to a limited number of speakers, we select for each claimed speaker a cohort set of three (same gender) whose acoustic models are closest to the claimed model. In the testing mode, each cohort speaker's 160 test tokens (10 utterances x 8 testing sessions x 2 repetitions) are tested against each claimed speakers' 10-word models.

To identify the fuzzy densities for the cohort speakers, we select the values of α by means of the training data. The range of α was specified to be from 1 to 50, and a unit step size was applied in the incremental trial process. It was observed that using different values of α for different speakers could give more reduction in the equal error rates. However, as an intial investigation we chose the same value for each gender set, that is α = 10 for the female cohort set and α = 1 for the male cohort set. As a result, Table 1 shows the mean equal-error rates for the 16 speakers with three codebook sizes of 32, 64 and 128 entries. The total average EER reductions by the fuzzy fusion (FF) in comparison with the geometric mean (GM) for the three codebook sizes of 32, 64 and 128 are (5.87-4.20)= 1.67%, (4.23-3.17)= 1.06%, (3.53-2.66)= 0.87%, respectively. Through these results, it can be seen that the speaker verification system using the fuzzy fusion is more favorable than using the geometric mean method.

Table 1. Equal error rates (%EERs) for the 16 speakers using geometric mean (GM) and fuzzy fusion (FF)

	GM			FF		
	Codebook Size			Codebook Size		
Speaker	32	64	128	32	64	128
f1	4.17	3.01	2.40	1.80	1.19	1.19
f2	5.98	1.19	1.79	1.19	0.60	1.20
f3	9.90	5.66	3.67	7.79	3.70	2.33
f4	0.00	0.00	0.00	0.00	0.00	0.00
f5	1.78	1.78	0.59	1.19	0.60	0.00
f6	6.67	3.01	1.80	2.41	0.59	0.00
f7	7.38	4.32	3.61	6.48	4.00	2.30
f8	12.76	9.73	9.22	10.05	8.22	7.62
m1	3.07	3.05	3.06	3.03	3.03	2.43
m2	4.17	1.28	1.22	3.14	1.22	1.22
m3	7.03	7.00	6.32	6.87	6.85	5.92
m4	10.77	8.28	7.90	8.29	6.89	6.91
m5	2.70	2.44	1.80	1.62	0.63	1.19
m6	8.43	7.44	6.53	7.53	5.47	4.72
m7	7.18	5.88	4.83	6.86	4.88	3.65
m8	1.83	3.01	2.40	1.80	1.21	1.19
Female	6.08	3.66	2.89	3.50	2.48	1.92
Male	5.65	4.80	4.17	4.89	3.86	3.40
Total	**5.87**	**4.23**	**3.53**	**4.20**	**3.17**	**2.66**

6. CONCLUSIONS

A fusion algorithm based on the fuzzy integral has been proposed and implemented in the similarity normalization for speaker verification. The the experimental results show that the application of the proposed method is superior to that of the conventional normalization. The key difference between the two methods is that the assumption of equal likelihood is not necessary for the fuzzy integral based normalization due to the concept of the fuzzy measure. One important issue arising here for further investigation is the *optimal* identification of the fuzzy densities in terms of the constant α, which can offer flexibility and have great effect in the fuzzy fusion. Similar work on this fuzzy fusion for numeral recognition was discussed in [13]. At present, the fuzzy densities were only determined based on a rough estimate of the values for α using a small range of integers.

7. REFERENCES

1. S. Furui, An overview of speaker recognition technology, *Proceedings of Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny (Switzerland), 1994, pages 1-9.

2. B.S. Atal, Effective of linear prediction characteristics of speech wave for automatic speaker identification and verification, *J. Acoust. Soc. Am.* **55**, 1304-1312 (1974).

3. S. Furui, Cepstral analysis techniques for automatic speaker verification, *IEEE Trans. Acoust. Speech Signal Processing* **29**, 254-272 (1981).

4. A.L. Higgins, L. Bahler and J. Porter, Speaker verification using randomnized phrase prompting, *Digital Signal Processing* **1**, 89-106 (1991).

5. T. Matsui and S. Furui Concatenated phoneme models for text variable speaker recognition, *Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Minneapolis (USA), 1993, pages 391-394.

6. G. Gravier and G. Chollet, Comparison of normalization techniques for speaker verification, *Proceedings of Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, Avignon (France), 1998, pages 97-100.

7. C.S. Liu, H.C. Wang and C.H. Lee, Speaker verification using normalization log-likelihood score, *IEEE Trans. Speech and Audio Processing* **4**, 56-60 (1996).

8. M. Sugeno, Fuzzy measures and fuzzy integrals – A survey, *Fuzzy Automata and Decision Processes*, M.M. Gupta, G.N Saridis and B.R. Gaines, Eds. Amsterdam: North-Holland, pp. 89-102, 1977.

9. K. Leszczynski, P. Penczek and W. Grochulski, Sugeno's fuzzy measure and fuzzy clustering, *Fuzzy Sets and Systems* **15**, 147-158 (1985).

10. T. Murofushi and M. Sugeno, An interpretation of fuzzy measure and the Choquet integral as an integral with respect to a fuzzy measure, *Fuzzy Sets and Systems* **29**, 201-227 (1989).

11. Y. Linde, A. Buzo and R.M. Gray, An algorithm for vector quantization, *IEEE Trans. Comm.* **28**, 84-95 (1980).

12. L.A. Zadeh, Fuzzy sets, *Information and Controls* **8**, 338-353 (1965).

13. Z. Chi, H. Yan and T. Pham, *Fuzzy Algorithms with Applications in Image Processing and Pattern Recognition*, (World Scientific, Singapore, 1996).