Fuzzy-Clustering-Based Decision Tree Approach for Large Population Speaker Identification

Yakun Hu, Dapeng Wu, and Antonio Nucci

***Abstract*—In this paper, we address the problem of large population speaker identification under noisy conditions. Major techniques for speaker identification is based on Mel-Frequency Cepstral Coefficients (MFCC), Gaussian Mixture Model (GMM) and Universal Background Model (UBM) which we call MFCC+GMM and MFCC+GMM+UBM. The approaches are known to perform very well for small population identification under low-noise conditions. However, the increase of population size can cause performance degradation of these schemes under noisy conditions. To mitigate this limitation, we propose a fuzzy-clustering-based decision tree approach. The key idea of our approach is to 1) use a decision tree to hierarchically partition the whole population into groups of small size, and determine which speaker group at the leaf node a speaker under test belongs to, and 2) apply MFCC+GMM to the selected speaker group for speaker identification. The advantage of our approach is that we use features that are independent from MFCC to partition speakers into groups and only apply MFCC+GMM to speaker groups at the leaf level. The key challenge in our design is how to achieve a low error probability of decision-tree-based classification. To address this, we adopt fuzzy clustering in constructing the tree for population partitioning, i.e., at each level, a speaker may belong to multiple groups. Such redundancy increases the probability of classifying a speaker under test into a correct group/node on the tree. Another novelty of this paper is that we use pitch and five vocal source features to construct a six-level decision tree. Experimental results demonstrate that our approach outperforms MFCC+GMM and MFCC+GMM+UBM with higher accuracy and lower complexity for large population identification under additive white Gaussian noise (AWGN) conditions.**

***Index Terms*—Large Population Speaker Identification, Hierarchical Decision Tree, Fuzzy Clustering, GMM, MFCC**

I. Introduction

Speaker identification [1] is an example of biometric system that has many useful applications. In speaker identification, given an input speech, the task is to determine the unknown speaker's identity by selecting one from the whole population of speakers registered in the system. In this paper, we consider large population speaker identification under noisy conditions. Specifically, there are a large number of registered speakers in our system and there is a mismatch between training and testing caused by noisy conditions (i.e., training samples are clean but testing samples are corrupted by additive noise).

Yakun Hu and Dapeng Wu are with Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611. Correspondence author: Prof. Dapeng Wu, wu@ece.ufl.edu, http://www.wu.ece.ufl.edu. Antonio Nucci is with Narus, Inc., 570 Maude Court, Sunnyvale, CA 94085.

The major technique for speaker identification is based on MFCC (Mel-Frequency Cepstral Coefficients) and GMM (Gaussian Mixture Model) [1]. Some important GMM-based approaches including the Universal Background Model (UBM) approach have been proposed [2], [3]. In this paper, we call them the MFCC+GMM approach and the MFCC+GMM+UBM approach. Another emerging technique which becomes very population is the i-vector approach (including the joint factor analysis approach) [4], [5]. The i-vector approach has been widely used for speaker verification. However, it seems not be directly applied to speaker identification yet. The i-vector approach usually requires a large number of data to perform well and the computational complexity can be high when applying i-vector to speaker identification especially for large population case. In our paper, we use the MFCC+GMM approach and the MFCC+GMM+UBM approach as the benchmarks for performance comparison.

The approaches based on MFCC and GMM are known to perform very well for small population speaker identification under low-noise conditions [1], [2]. However, they also have some drawbacks. The first drawback is that they suffer from the mismatch between training and testing caused by noisy conditions. The noisy conditions can severely degrade the identification performance. The second drawback is actually a common problem of almost all existing speaker identification techniques. The success of almost all existing identification systems (including GMM-based systems) lies in the fact that they are trained on datasets with only a relatively small population. However, it is pretty straightforward that when the population has a significant increase (e.g., thousands of registered speakers or even more), the probability of identification errors will significantly increase, accordingly. Unfortunately, there are not much existing research work studying this problem. Some papers mainly focused on reducing the computational complexity in large population cases at the cost of a very slight accuracy loss [6]–[8]. In some other papers which claimed to deal with large population identification, the experiments were actually carried out on datasets with only hundreds of registered speakers [9], [10]. In [11], Chaudhari et al. attempted to address the truly large population identification problem and they proposed a derivative of MFCC+GMM and achieved a good accuracy on the IBM internal dataset consisting of 10013 speakers. The experiments in [11] were conducted when training and testing conditions are matched without additive noise or channel variations. Nevertheless, the population becomes an extremely important impact factor of the identification performance

under noisy conditions. Some existing research have provided evidences to support this conclusion. One evidence is from [12] in which Reynolds showed the accuracy of MFCC+GMM as a function of the population size on NTIMIT database which contains speech samples degraded by noise and bandlimiting [12]. Experimental results shown in [12] indicate that the identification accuracy steadily decreases as the population size increases and the largest drop in accuracy occurs when the population size increases to 100. With the full 630 speaker population, there is about 30% loss in accuracy compared with 10 speaker population case. Another evidence comes from our own experimental results. Fig.1 shows the accuracy v.s. population for MFCC+GMM on our own speech dataset (the specific description will be given in Section V) in the scenario of additive white Gaussian noise (AWGN) with a 30dB signal-to-noise ratio (SNR). From the figure, we also can see there is a steady and significant accuracy loss as the population goes up. As a conclusion, approaches based on MFCC and GMM have achieved great success in speaker identification. However, two factors including the additive noise that is ubiquitous in the environment and the practical requirements of large population identification are greatly limiting the approaches in real applications.

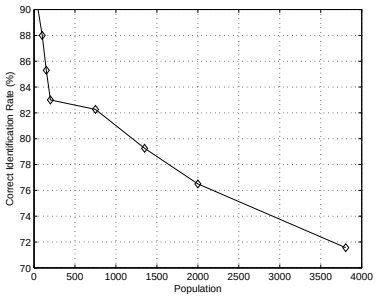


Fig. 1. Accuracy v.s. Population for MFCC+GMM.

To mitigate the limitations of MFCC+GMM (including MFCC+GMM+UBM) and improve the performance of large population speaker identification under noisy conditions, we proposed a fuzzy-clustering-based hierarchical decision tree approach. The key idea of our approach is that we use a decision tree to partition the original large population into subgroups in a hierarchical way. For a speaker under test, we first conduct the decision-tree-based classification (i.e., determine which speaker group at the leaf node of the tree the speaker belongs to) and then apply the MFCC+GMM identification approach to the selected speaker group at the leaf node to determine the speaker identity. The decision tree has multiple levels and the population partitioning is conducted from upper levels of the tree to its lower levels. The root node of the tree represents the universal set containing all registered speakers. At each level of the tree, we use a speech feature to do speaker clustering, i.e., a node (or a speaker group) splits into several child nodes (or speaker subgroups) at its lower level. In this process, speakers with similar speech feature are put into a same child node whereas speakers with dissimilar speech feature are put into different child nodes. Then, each child node contains a smaller population size than its parent node. Thus, at the bottom level, each speaker group at the leaf node has a very small population size and the population reduction is achieved. When the hierarchical decision tree is constructed, for a speaker under test, we conduct the decision-tree-based classification from the top of the tree to its bottom. At each level, we determine which speaker group or node the speaker belongs to. At the bottom level, we select one and only one speaker group at the leaf node that the speaker belongs to and apply MFCC+GMM to the selected speaker group for speaker identification. The advantage of our approach is that 1) we only apply MFCC+GMM to the speaker group at the leaf node with a very small population size instead of applying it to the original large population, and 2) since we only use speech features that are independent from MFCC to cluster speakers into groups, speakers with similar MFCC may not be put into a same speaker group and the probability of speakers having similar MFCC is much lower in the speaker group at the leaf node than at the root node containing the whole population. Hence MFCC+GMM can perform well in the speaker group at the leaf node with a much higher correct identification rate as well as a much lower computational complexity.

In the process of decision-tree-based classification, a speaker may be classified into an incorrect node or speaker group. What is worse, a classification error at any level will propagate through the tree and finally accumulates at the bottom level, i.e., if a speaker under test is classified into an incorrect node (or speaker group) at a level, the speaker will finally be classified into an incorrect speaker group at the leaf node. Then, there is no chance for us to correctly identify the speaker. Therefore, the key challenge in our design is how to achieve a low probability of classification error in the process of decision-tree-based classification for a speaker under test. We can use the conventional hard clustering in constructing the decision tree, i.e., a speaker only belongs to one node (or speaker group) at each level of the decision tree. However, it seems that the classification accuracy in the process of decision-tree-based classification could not be satisfactory in this case. For example, for those speakers on the boundaries between different groups, an inevitable feature deviation in the testing phase, even very small, will almost guarantee an error of classification. Also, for speakers whose feature values have relatively large deviation from sample to sample, it is difficult to prevent from classifying them into the incorrect speaker groups. To achieve a satisfactory performance of decision-tree-based classification, we proposed to adopt fuzzy clustering in constructing the decision tree, i.e., a speaker may belong to multiple nodes (or speaker groups) at each level of the decision tree. For a speaker under test, such redundancy introduced by fuzzy clustering can greatly increase the probability of the speaker being "captured" by a correct node (or speaker group) at each level of the tree. Thus, the probability of classifying a speaker under test into a correct speaker group at the leaf node can greatly increase, accordingly.

Another novelty of this paper is that we derived six features (including pitch and five vocal source characteristics) to construct a six-level decision tree. The six features are believed to 1) be able to discriminate different groups of

speakers, 2) be independent from MFCC, 3) be independent from each other, and 4) be robust to additive noise (e.g., AWGN). We evaluate the performance of the six-level decision tree and compare the identification performance with the MFCC+GM approach and the MFCC+GMM+UBM approach on our own dataset in the scenario of AWGN. Experimental results indicate that our approach outperforms the MFCC+GMM approach and the MFCC+GMM+UBM approach with higher correct identification rate (e.g., 8% increase compared with MFCC+GMM+UBM for 30dB SNR and 30s testing length) and lower computational complexity which meets the requirement of real-time applications.

The remainder of this paper is organized as follows. In Section II, we propose the fuzzy-clustering-based hierarchical decision tree. Section III specifically describes the speech features used in the decision tree for speaker clustering. In Section IV, we present the fuzzy clustering algorithm adopted in the decision tree. Section V shows the experimental results and Section VI concludes the paper.

II. Fuzzy-Clustering-Based Hierarchical Decision Tree

In this section, we specifically describe our fuzzy-clustering-based hierarchical decision tree for speaker identification. In Section II-A, the system diagrams using the hierarchical decision tree are shown. Section II-B explains the design of the decision tree approach.

A. Diagrams of Fuzzy-Clustering-Based Hierarchical Decision Tree

There are two units in our identification system using fuzzy-clustering-based hierarchical decision tree: decision-tree-based classification and speaker identification at leaf nodes of the tree. Decision-tree-based classification has two phases: training phase (i.e., speaker clustering for the construction of the hierarchical decision tree) and testing phase (i.e., the process of determining which speaker group at the leaf node a speaker under test belongs to). Speaker identification at leaf nodes also consists of training phase and testing phase.

Fig.2 shows the construction of the hierarchical decision tree. In the figure, $C_{n_1,n_2,\cdots,n_L,n_{L+1}}$ denotes the $n_{L+1}th$ node representing a speaker group at level L, where $L = 0, 1, \cdots$. Every node is a set of speakers. The root node C_1 at level 0 represents a single speaker group containing a total of N speakers, where N has a large value (e.g., >1000) in our problem. The parent node of $C_{n_1,n_2,\cdots,n_L,n_{L+1}}$ at level L is $C_{n_1,n_2,\cdots,n_L}$ at level $L-1$. At each level, the speaker clustering is conducted, i.e., a parent node at an upper level is split into several child nodes at its lower level. The speaker clustering is carried out hierarchically from upper levels to lower levels and is completed until each leaf node of the tree can be labeled by a particular speaker group which contains a sufficiently small number of speakers (e.g., <50). As mentioned in Section I, we adopt fuzzy clustering at each level of the tree. Specifically, a speaker in a parent node at a level may belong to more than one child nodes at its lower level. The fuzzy clustering will be specified in Section IV.

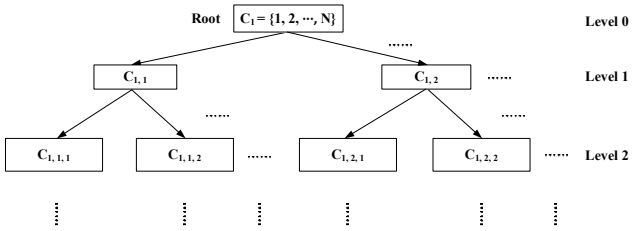
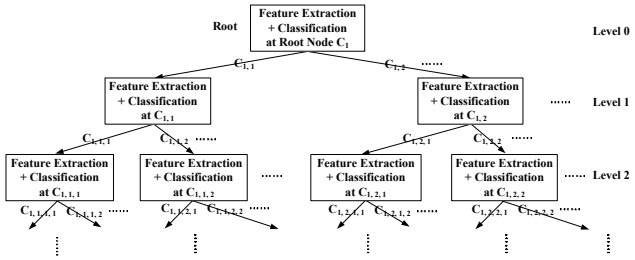
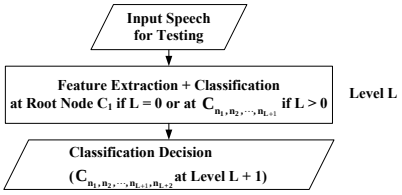


Fig. 2. Construction of Hierarchical Decision Tree.
(Fuzzy Clustering: $\bigcup_{n_{L+1}} C_{n_1,n_2,\cdots,n_L,n_{L+1}} = C_{n_1,n_2,\cdots,n_L}$ and $|C_{n_1,n_2,\cdots,n_L,p} \bigcap C_{n_1,n_2,\cdots,n_L,q}| \geq 0$, where $p \neq q$, $L = 1, 2 \cdots$ and $|\cdot|$ denotes the cardinality of a set)

Fig.3 illustrates the classification based on hierarchical decision tree, where Fig.3(a) gives the overall structure of the hierarchical decision tree for classification and Fig.3(b) specifies the input and output of a node at each level of the tree for classification. As indicated from Fig.3(a), for a speaker under test, the decision-tree-based classification proceeds from the top of the tree to its bottom. At each level of the tree, Fig.3(b) shows that the input of a node is an input speech for testing. After feature extraction and classification procedure, the node outputs the classification decision, i.e., which speaker group at the child node at the lower level a speaker under test is determined to belong to. At each level, one and only one node is enabled and thus, for an input testing speech, an unique path from the root node of the tree to one leaf node is enabled.



(a) Overall Structure of Hierarchical Decision Tree for Classification



(b) Input and Output of a Node at Level L

Fig. 3. Classification based on Hierarchical Decision Tree.

When a leaf node is enabled, the decision-tree-based classification will be terminated and MFCC+GMM will be applied to the speaker group at the enabled leaf node for speaker identification. Since we only use those speech features that are independent from MFCC to conduct speaker clustering via decision tree, the probability of having speakers with similar MFCC is significantly reduced in the speaker group at the leaf node. Thus, MFCC+GMM can perform well. Notice that the speaker group at the enabled leaf node contains a small population, we do not need a GMM with a large number of

mixtures to complete the identification task and this further ensures the computational complexity to be low.

B. Why Hierarchical Decision Tree?

In large population speaker identification, why it's feasible to use hierarchical decision tree for population reduction? This is because human speech does contain many useful features that can be used to cluster speakers into groups. Speaker groups do exist that speakers sharing with a similar speech feature are in a same group whereas speakers having different speech features are from different groups. For example, speakers with different genders can be distinguished by using pitch feature [13]; based on different movement patterns of the vocal cords during utterances, different speaker groups could be obtained; Many emerging features which are independent from MFCC may indicate different speaker groups [14]. In summary, human speech has many different attributes and it's feasible to cluster speaker into groups by using various speech features. At each level of our hierarchical decision tree, we try to find different speaker groups by examining a certain attribute of speech.

In our hierarchical decision tree, the speaker clustering is carried out orderly and independently from the upper levels to the lower levels. There is another alternative way that we can jointly use all features to complete the clustering at one time. Why we adopt the hierarchical way? There are basically two reasons. On one hand, to classify a speaker under test into a cluster (or speaker group) where MFCC+GMM is used for speaker identification, our approach requires lower computational complexity. Our approach uses low-dimensional feature data (e.g., one-dimensional) and the classification is much less complicated than the one using high-dimensional feature data. Moreover, let us do a complexity analysis of the classification. Suppose eight clusters are created by using each feature, then jointly using M features will result in 8^M clusters and need a computational complexity of $O(8^M)$ during the testing phase. In contrast, a hierarchical tree constructed by M features will also result in 8^M clusters at the leaf level but only incurs a computation complexity of $O(M)$ during the testing phase. On the other hand, as will be shown in Section III, since features used at different levels of the tree are required to be independent from each other, the classification performance of our approach and the one jointly using multiple features should be close to each other.

Some researchers proposed to combine MFCC and the features that are complementary to MFCC for speaker identification. Ezzaidi et al. put forward to combine pitch and MFCC for speaker identification [15]. In [16], Wang and Zheng integrated wavelet octave coefficients of residues (WOCOR) into MFCC for speaker identification. In [17], Hosseinzadeh et al. derived a set of spectral features from the excitation component of speech and combined them with MFCC for speaker identification by using GMM. Nakagawa et al. proposed to combine MFCC and phase information for speaker identification [18]. Those approaches use different fusion techniques to combine likelihood scores based on different features for speaker identification. Although fusion techniques are fairly mature in speaker recognition, some additional training is required to obtain the optimal weights or parameters for the feature combination [19]. For large population speaker identification, the key drawback of those approaches is that they require high computational complexity and they are not sufficiently scalable. From one aspect, all those approaches need to combine scores for different features (including MFCC) against all speaker models and thus are not applicable to large population identification due to the unacceptably high complexity. From another aspect, when a new feature is available, the fusion approach may need to be redeveloped in order to accommodate the new feature and this greatly reduces the scalability of the approaches. Comparatively, in our approach, we believe that those complementary features to MFCC only can provide a certain profile of the speech and thus are more suitable for grouping speakers rather than identifying speakers directly. We derived some complementary features and used them to construct a decision tree for classifying speakers into subgroups before using MFCC for identification. In this way, the complexity can be significantly reduced for large population identification since GMM likelihood scores are only calculated against a small number of speaker models. In our decision tree, the classification at each level and the identification at the leaf node are independently conducted. Therefore, it's not difficult to accommodate a new feature by just adding one more level to the existing tree without having any effect on the original design.

III. Speech Features for Speaker Clustering

To achieve good performance, features used in our approach for clustering should meet the following requirements: 1) a good feature should be very capable of discriminating different groups of speakers; 2) features used at different levels of the tree should be independent from each other; 3) all features should be independent from MFCC used at the leaf node for identification; 4) all features should be robust to additive noise. This section will describe the six features we derived.

A. Feature Description

All features we used fall into the category of vocal source feature. The source-filter model of speech production [20] tells us that speech is generated by a sound source (i.e., the vibration of vocal cords) going through a linear acoustic filter (i.e., the combination of the vocal tract and the lip). MFCC mainly represents the vocal tract information. The vocal source is believed to be an independent component from the vocal tract and is able to provide some speaker-specific information. This is why we are interested in extracting vocal source features for speaker clustering.

The first feature we derived is pitch or fundamental frequency. The pitch period is the period of the vocal source vibration and can be estimated from the period of a voiced sound. The rest of five features are all related to the vocal source excitation of voiced sounds. We extract them from the linear predictive (LP) residual signal [21] which is a good, though not exact, representative of the vocal source excitation.

It is well known that a LP residual signal of a voiced sound is virtually periodic and in each cycle, there are both a positive pulse and a negative pulse. For different speakers, the shape of the pulses are very different. The five vocal source features we use are width of the positive pulse, skewness of the positive pulse, skewness of the negative pulse, PAR of the positive pulse within one cycle and PAR of the negative pulse within one cycle. Notice that we do not use the width of the negative pulse as a feature since experimental results indicate that it does not perform as well as the other five. The related discussion will be given in Section V-A2.

B. Feature Extraction

In this section, we will specify how the six features are extracted from the speech signal.

1) Pitch Extraction: In our work, YIN algorithm [22] is used for pitch feature extraction. Fig.4 shows the input and the output of the pitch extraction module using YIN algorithm. As indicated from the figure, given a continuous speech as the input, the module first decomposed it into N_F frames. The frame length is 25ms and the frame shift length is 10ms. For the i-th frame ($i = 1, 2, \cdots, N_F$), we obtain the pitch estimation p_i and the probability of the frame being voiced denoted as Pr_i. Since the reasonable pitch range of human speech is from 50Hz to 550Hz [23], we drop all pitch estimations which are lower than 50Hz or higher than 550Hz. We also discard all pitch estimations that are extracted from frames whose probability of being voiced are below 0.8. By doing so, we can remove all potential outliers and obtain a set of reliable pitch estimations.

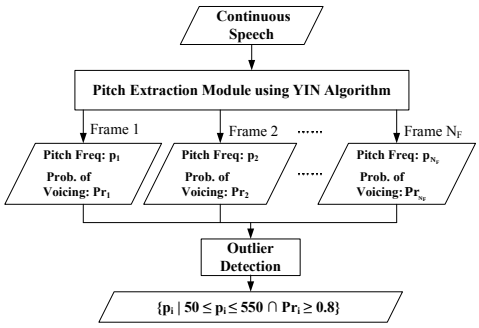


Fig. 4. Pitch Feature Extraction.

2) Vocal Source Features Extraction: We developed our own algorithm to extract all five vocal source features. As similar as pitch extraction, the vocal source features are only derived from voiced speech frames. Fig.5 shows the process of vocal source feature extraction. Given a continuous speech as the input, it is decomposed into short-time frames as similarly as shown in Section III-B1. For each speech frame, we determine whether it's voiced or not based on the energy and the zero crossing rate of the frame. If it is voiced, we apply the well-known Levinson-Durbin algorithm to the frame to estimate the LP coefficients. As the sampling rate is 11.025kHz, the LP order is set to be 14. This is reasonable because a rule of thumb to choose the LP order is to use 1 complex pole per kHz plus 2-4 poles to model the radiation and glottal effects [20]. By using those LP coefficients, we obtain the LP residual signal and extract all five vocal source features from the LP residual signal.

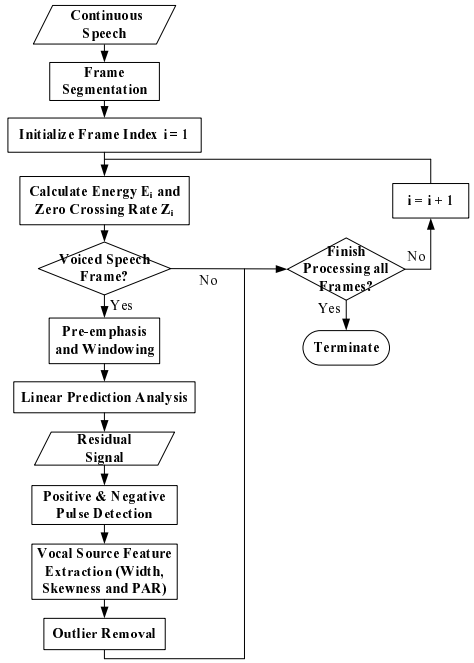


Fig. 5. Vocal Source Features Extraction.

The vocal source feature extraction is not specified here since it is not the focus of this paper. The last step of the vocal source features extraction is the outlier removal just like the pitch feature extraction. Since there is no solid knowledge about the reasonable range of the vocal source feature values, we can not set lower and upper limits as we did for pitch feature but we can do it by statistical analysis. It is observed that, for each of five vocal source features, feature estimations roughly have a normal distribution. A certain percentage of the feature estimations with highest values and a same percentage of the feature estimations with lowest values are treated as outliers and should be removed. The selection of the percentage for different vocal source features will be shown in Section V.

C. Feature Evaluation

We evaluate our features to see whether they meet the requirements mentioned at the beginning of this section.

- All six features fall into the category of vocal source features which characterize some attributes of the movement of the vocal cords. As the movement of the vocal cords is related with the glottis structure and the speaking habit of a speaker during utterances, it should be pretty stable for a specific speaker and be quite different from different speakers. Thus, all six features are capable of discriminating different groups of speakers.
- All six features characterize totally different attributes of the movement of the vocal source. Pitch represents the period of the vocal cords movement; width features

measure how long time the vocal cords are in open state within one movement cycle; skewness features provide information about the rate of vocal cords opening and closing; PAR features characterize how the energy distributes in the process of vocal cords movement. Therefore, the six features should be independent from each other and can be used at different levels of the tree to give a significant population reduction in total.

- All six features are vocal source features and are independent from MFCC which characterizes the vocal tract during utterances. Therefore, they should be independent from MFCC used at leaf nodes of the tree.
- All six features are robust to additive noise (e.g., AWGN). There are mainly two reasons. On one hand, all features are only extracted from voiced speech frames which usually have relatively high SNR. On the other hand, all features do not learn details of speech waveforms but are some quantities that characterize some attributes of the vocal source in a generalized way and accordingly, are more robust to additive noise. For example, in the pulse portion of the LP residual signal extracted from a voiced frame, SNR is relatively high and all the five vocal source features can be very insusceptible to additive noise; additive noise can distort the specific speech waveform but the periodicity of the speech signal can probably be preserved and pitch can still be well extracted.

The analysis above indicates that all six features meet the requirements and can be used to construct our hierarchical decision tree for speaker clustering.

IV. Fuzzy Clustering

In this section, we will present the fuzzy clustering algorithm used in our decision tree.

A. Why Fuzzy Clustering?

The decision-tree-based classification performance is crucial to the success of our approach for large population speaker identification under noisy conditions. The target is that we want to achieve a high accuracy of decision-tree-based classification with a significant population reduction in the speaker groups at leaf nodes. In terms of classification accuracy, suppose we have a total of L_T speech features for speaker clustering and we can construct an L_T-level tree by using one feature at one level. Denote the correct classification rate achieved by one-level decision tree using feature l ($l = 1, 2, \cdots, L_T$) for speaker clustering as Acc_l and Acc_l is defined as the probability of classifying a speaker under test into a correct speaker group at the leaf node of the one-level tree. When we construct an L_T-level tree by using feature l at level l, we denote the overall accuracy of decision-tree-based classification as Acc and Acc is defined as the probability of classifying a speaker under test into a correct speaker group at the leaf node of the L_T-level tree. Then, it's straightforward to obtain

$$Acc \approx Acc_1 Acc_2 \cdots Acc_{L_T} \tag{1}$$

If we have six features and each feature can achieve a 97% correct classification rate, then we can do a simple calculation that for the six-level tree, the overall accuracy is only $(97\%)^6 = 83.3\%$. The degradation of the classification accuracy is remarkable when the number of levels increases. Hence, at each level of the tree, we must ensure a sufficiently low probability of classification error so that the overall accuracy of a multi-level decision tree can be satisfactory.

In respect of population reduction, for a hierarchical decision tree, we can define its population reduction rate as 100% minus the ratio of the population averaged over all leaf nodes of the tree to the original whole population. The population reduction rate tells you how many percent the population is reduced after speaker clustering via the decision tree and a higher rate means a more population reduction achieved. Suppose we construct an L_T-level decision tree and let PR_{L_T} denote the population reduction rate achieved by the tree. If the total number of registered speakers is N, then the population averaged over all leaf nodes denoted as N_{leaf} can be obtained by

$$N_{leaf} = N(1 - PR_{L_T}) \tag{2}$$

If we only use speech features that are independent from MFCC to construct the decision tree for speaker clustering, then the overall correct identification rate achieved by our speaker identification system using the L_T-level decision tree can be approximated as

$$CIR \approx Acc \times A_G(N_{leaf}) \tag{3}$$

where A_G is the correct identification rate achieved by MFCC+GMM and it's a function of the population size. From Equation (3), we know that in order to achieve better identification performance, we not only want the accuracy of decision-tree-based classification Acc to be higher but also want the leaf nodes to contain a population size as small as possible because we have shown that A_G will decrease as population increases in Section I. However, generally, a higher population reduction rate results in a lower classification accuracy. This is because when the clustering algorithm tries to load a smaller population into each speaker group or node at a certain level, the dynamic range of feature values of speakers in each group becomes smaller and the "distance" between neighbouring speaker groups also become smaller. In the process of decision-tree-based classification, an inevitable feature inconsistency between training and testing will cause an incorrect classification with a higher probability.

To achieve a low error probability of decision-tree-based classification, we propose to use fuzzy clustering at each level of our decision tree. Fuzzy clustering, which is different from the conventional hard clustering, is a class of algorithms for cluster analysis that allow the objects to belong to several clusters simultaneously, with different degrees of membership [24]. Different from the common fuzzy clustering algorithms, at each level of the tree, our fuzzy clustering algorithm allows one speaker to belong to multiple speaker groups or nodes, simultaneously. That is to say, our algorithm does not assign degrees of membership or alternatively, for each speaker group

at a level of the tree, the degree of membership is either 0 or 1 but there may be multiple speaker groups whose degrees of membership are 1. At each level of the tree, the classification error mainly comes from those speakers on the boundaries between different speaker groups or those speakers who have relatively low feature stability. When conducting speaker clustering, if we put those speakers into all speaker groups that they may belong to, then in the process of decision-tree-based classification, those speakers can be "captured" with a much higher probability by correct speaker groups in which they can be found. The probability of classification error can be significantly reduced, accordingly. A performance comparison between hard clustering and fuzzy clustering will be made in Section V-A2. One thing needs to be pointed out that fuzzy clustering results in a less population reduction at leaf nodes due to the redundancy introduced among speaker groups at each level. More redundancy leads to a higher classification accuracy but a less population reduction so there is a trade-off. Our fuzzy clustering algorithm aims at introducing the most appropriate redundancy and achieving a satisfactory classification accuracy with a population reduction as much as possible.

B. Fuzzy Clustering Algorithm

Fig.6 shows the flow chart of the fuzzy clustering algorithm used at level L of the decision tree. The algorithm applies to every feature we derived so that the flow chart does not specify the feature. As shown in the figure, we cluster all speakers belonging to $C_{n_1,n_2,\cdots,n_L}$ at level $L-1$ into several child nodes at level L. To ensure that there is a sufficient number of speakers for clustering, before conducting any clustering operation, we first count the number of speakers contained in $C_{n_1,n_2,\cdots,n_L}$ at level $L-1$. If the number of speakers is less than a pre-determined number (e.g., 20), no clustering operation will be carried out. Since the population contained in $C_{n_1,n_2,\cdots,n_L}$ is small enough for MFCC+GMM to yield a satisfactory performance of speaker identification, $C_{n_1,n_2,\cdots,n_L}$ will be treated as a leaf node to which MFCC+GMM will be applied for speaker identification. If there is a sufficient number of speakers in $C_{n_1,n_2,\cdots,n_L}$ at level $L-1$, the fuzzy clustering at level L will be conducted. We first do feature extraction and obtain the feature denoted as $F_{i,j}$. Here, i ($i \in C_{n_1,n_2,\cdots,n_L}$) is the speaker index and j ($j = 1, 2, \cdots, N_i$) is the feature index, where N_i denotes the total number of feature estimations of speaker i. Notice that for each of the six features we derived, the feature data is one-dimensional. Then, instead of using the raw feature data, we use the statistics of feature data for clustering. Specifically, for speaker $i \in C_{n_1,n_2,\cdots,n_L}$, we first calculate the mean and the standard deviation of the feature data as follows:

$$\mu_i = \frac{\sum_{j=1}^{N_i} F_{i,j}}{N_i} \tag{4}$$

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{N_i}(F_{i,j}-\mu_i)^2}{N_i - 1}} \tag{5}$$

Then, a confidence interval $[\mu_i - \lambda\sigma_i \ \mu_i + \lambda\sigma_i]$ is constructed for speaker i, where λ is a pre-determined coefficient. For speaker i, the two new statistical data $\mu_i \pm \lambda\sigma_i$ can be a good statistical representation of the raw feature data. Next, let $D = \{\mu_i - \lambda\sigma_i, \mu_i + \lambda\sigma_i\}$ which is a data set containing the two statistical data of all speakers belonging to $C_{n_1,n_2,\cdots,n_L}$. D is fed into Lloyd's algorithm and a partition vector $[P_0, P_1, \cdots, P_M]$ is output, where M is the total number of clusters adopted by Lloyd's algorithm. Finally, based on the partition vector, we create all M clusters (i.e., speaker groups). For speaker i and cluster m ($m = 1, 2, \cdots, M$), if $[\mu_i - \lambda\sigma_i \ \mu_i + \lambda\sigma_i] \bigcap [P_{m-1} \ P_m] \neq \emptyset$, we let $i \in C_{n_1,n_2,\cdots,n_L,m}$. By doing so, we select all those clusters which there is a probability that a speaker belongs to and replicate the speaker into all these selected clusters.

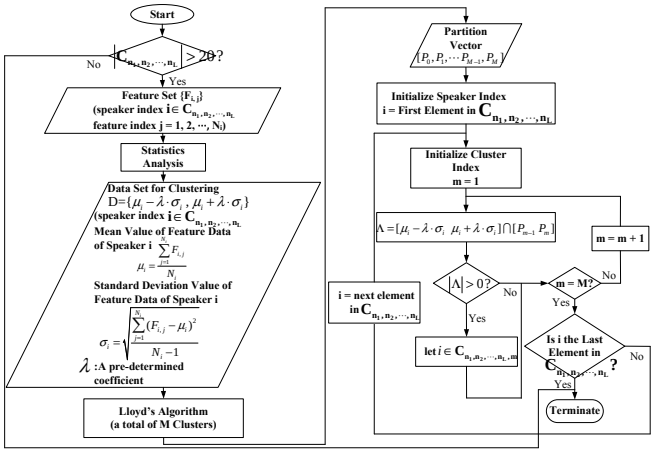


Fig. 6. Flow Chart of Fuzzy Clustering Algorithm at Level L of the Decision Tree.

From the description above, we know that our algorithm consists of two parts: partition vector determination and replication. In the aspect of partition vector determination, some statistical data rather than the raw feature data are used to determine the cluster boundaries. There are at least two benefits to do so. One is that the computational complexity of using statistical data is much lower, compared with the case of using all raw feature data, especially for large population identification. The other one, also more important, is that it prevents from using those feature estimations that are out of the confidence interval for analysis since those "abnormal" feature estimations may give rise to an inaccurate clustering result. In this sense, it acts like a further outlier removal that we use statistical data. In the aspect of replication, for each speaker, we only use those reliable feature estimations that fall into the confidence interval to determine the replication and it can make the replication reasonable. On one side, those feature data that are out of the confidence interval only occur with a very low probability. If we use those data for replication, then it will induce an over-replication and thus a less population reduction will be achieved. On the other side, the feature data that are out of the confidence interval may not well "represent" the speaker. The replications based on these feature data are probably useless because in the decision-tree-based

classification, the probability of a speaker being classified into a cluster is very low if the speaker is replicated into this cluster based on those unreliable feature data.

In our fuzzy clustering algorithm, λ is an important parameter in constructing the confidence interval and conducting the replication. λ can tradeoff between the accuracy of decision-tree-based classification and the population reduction achieved by the decision tree. A higher value of λ gives rise to more replication which results in a higher classification accuracy but a less population reduction. Contrarily, a smaller value of λ brings about more population reduction but a lower classification accuracy. The selection of λ for different features will be shown in Section V-A2.

When the hierarchical decision tree is constructed by using the fuzzy clustering algorithm, for a speaker under test, we first determine which speaker group at the leaf node the speaker belongs to. Fig.7 shows the decision-tree-based classification conducted at level L. As shown in the figure, when the decision-tree-based classification is completed at level $L-1$ of the tree and $C_{n_1,n_2,\cdots,n_L}$ at level $L-1$ is enabled, we first determine whether $C_{n_1,n_2,\cdots,n_L}$ is a leaf node of the tree. If it is a leaf node, the decision-tree-based classification will be terminated and the MFCC+GMM identification approach will be applied to $C_{n_1,n_2,\cdots,n_L}$. If not, the decision-tree-based classification will be continuously conducted at level L. After feature extraction and outlier removal, a set of feature data is first obtained for the speaker under test. We then take the mean value of the feature data and make the classification decision at level L by comparing the mean value with the partition vector obtained by the Lloyd's algorithm in the fuzzy clustering. At last, based on the comparison, one and only one node at level L is enabled and the decision-tree-based classification at level L terminates. The classification will proceed from the enabled node at level L in the same way until one leaf node is finally enabled.

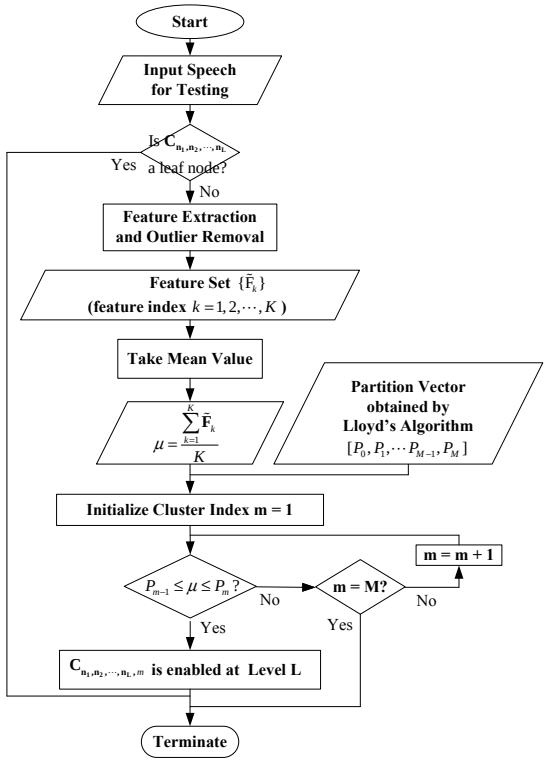


Fig. 7. Flow Chart of Decision-Tree-based Classification at level L.

V. Experimental Results

In this section, we implement the fuzzy-clustering-based hierarchical decision tree for large population speaker identification under noisy conditions and compare the performance with the MFCC+GMM+UBM approach and the MFCC+GMM approach used as the baseline systems. In Section V-A, we use the six features we derived for speaker clustering to construct a six-level hierarchical decision tree and evaluate its performance. Section V-B compares the performance of our fuzzy-clustering-based hierarchical decision tree approach, the MFCC+GMM approach and the MFCC+GMM+UBM approach for large population speaker identification under noisy conditions.

A. Performance Evaluation of Six-level Decision Tree

In this section, we use six features including pitch and five vocal source characteristics to construct a six-level hierarchical decision tree with appropriate parameters and evaluate the performance of the six-level decision tree including the classification accuracy and the population reduction rate. Section V-A1 describes the dataset used for experiments and the basic experimental settings. In section V-A2, we show the parameters used to construct the six-level tree and demonstrates the performance achieved by the six-level tree. We also compare the classification performance of the hard-clustering-based decision tree and the fuzzy-clustering-based decision tree.

1) Experimental Settings for Performance Evaluation of Six-level Tree: The dataset we use for the experiments is collected from the websites of online audiobooks such as www.audible.com, www.theaudiobookstore.com, etc. The online audiobooks cover a variety of topics such as business, history, science, etc. and each audiobook was recorded by a narrator. There are a large number of high-quality and clean audio samples recorded in English. For each narrator, minutes-long audio samples in mp3 format can be downloaded and all mp3 samples were converted into a 1-channel wav. format with a sampling rate of 11.025kHz. In this way, we create our own dataset which consists of totally 3805 speakers for the experiments. The dataset meets the requirement of large population speaker identification.

In the experiments, for each speaker, the training sample and the testing sample were recorded by a same microphone and the type of speech is read utterances. The amount of both training sample and testing sample are 30 seconds. Our experiments are conducted in the AWGN scenario. AWGN is commonly encountered in voice over IP (VoIP) environments and it's difficult to be perfectly eliminated from the desired speech signal. The white Gaussian noise with pre-determined energy is generated and is added to the clean testing samples by computer to meet the different requirements of SNR values.

2) Performance of Six-level Tree: In this section, experiments are conducted to determine the optimal parameters in constructing the six-level tree and to evaluate the performance of the tree. All experiments are conducted on the dataset described in Section V-A1.

For our decision tree approach, we know that adding a feature to an existing tree will bring in more population reduction as well as some loss of the classification accuracy. Therefore, we must make sure that the more population reduction brought by adding a feature will give us benefit even with some degradation of the classification accuracy. Usually, if the leaf node of an existing tree still contains a relatively large population (e.g., hundreds of speakers), then we believe that there is still some space for us to add more good features to the tree for further performance improvement. However, when the leaf node contains a relatively small population (e.g., dozens of speakers), then the population is too small to be further clustered and adding more features may not bring much performance improvement. What is worse, it may even give rise to performance degradation since it does not make any sense to cluster a small population and the performance is unpredictable. In a word, the number of levels of the tree depends on 1) the original population of identification problem and 2) the performance of the features used to construct the decision tree. If the original population is large, then we probably should construct a tree with more levels. If a good feature is available, then it should be able to bring more benefits by adding it to an existing tree. In our experiments, we consider a feature to be good if it can achieve a sufficiently high classification accuracy (e.g., $>$99%) as well as a good amount of population reduction (e.g., $>$40% reduction). By this criterion, the six features we developed are all good features. Since the population size we deal with is large (about 4000 speakers), therefore we use them all to construct a six-level tree. In Section III-A, we mentioned that the width of the negative pulse is also a candidate of vocal source features. However, it does not perform sufficiently well so that it may not give us benefit by adding it to the tree. Therefore, we discard it in stead of using it to construct a seven-level tree. If there are other good features available, we should add them to the existing six-level tree for further performance improvement.

In our decision tree approach, the parameters of pitch feature include the value of λ to construct the confidence interval and the number of clusters adopted by the Lloyd's algorithm; for the other five vocal source features, besides the two parameters listed above, there is one additional percentage parameter regarding the outlier removal. Table I gives an example of our decision tree used in AWGN scenario with 25dB SNR. It shows the parameters of the six features used at different levels of the tree.

From the table, we can see that from the first level to the sixth level, the features used are pitch feature, PAR of the positive pulse, skewness of the positive pulse, PAR of the negative pulse, skewness of the negative pulse and width of the positive pulse, successively. Since all features are almost independent from each other, ideally, there should only be a little effect on the performance if the tree is constructed by using features in a different order. However, we do have a criterion that the features which are able to yield a better classification performance (i.e., a higher classification accuracy as well as more population reduction) are used in the upper levels. For example, pitch feature has the best classification performance among the six features so that it is used in the first level of the tree. The reason is that we always want to use a better feature to partition the population when the population is relatively large. If the population of a node is relatively small, then the performance of clustering a small population into groups is not predictable since subgroups may not exist in the population. In our fuzzy clustering algorithm, a speaker will probably be replicated into every subgroup no matter how good the feature is. Therefore, to make full use of every feature and to achieve the best classification performance, we always use better features in the upper levels and avoid applying good features to cluster small population into groups. In practice, when constructing our decision tree, we first use our dataset described in Section V-A1 to evaluate the classification performance of the six features and determine the order of the six features used in the tree, accordingly. This is how we figure out the order shown in Table I and this order will be adopted to construct our decision tree in all experiments. One thing needs to be pointed out that in the ideal case when each node of the tree always contain a sufficient number of speakers to be clustered, we believe the order should not make a difference because they are almost independent from each other.

The parameters shown in the table are selected by trying different combinations and are able to achieve a sufficiently high classification accuracy as well as a high population reduction rate. The classification accuracy and the population reduction rate achieved at different levels of the six-level decision tree are shown in the table. All 3805 speakers are used to do the classification test through the decision tree and the classification accuracy achieved at a level is calculated as the percentage of speakers being classified into a correct node at that level. The calculation of the population reduction rate is based on the definition in Section IV-A but we use a weighted average over all nodes at different levels of the tree. Specifically, to calculate the population reduction rate achieved at a level, we assign a weight to a node at that level and the weight is determined as the percentage of speakers being correctly classified into the node at that level. For instance, a total of 1000 speakers are correctly classified at a level of the tree. Among all, 100 speakers are correctly classified into a node at that level with a population size of 200. Then, when we calculate the weighted average population over all nodes at that level, the weight assigned to that node is $100/1000 = 10\%$ because 10% of all speakers are classified into a speaker group with a population size of 200. The weighted average population is reasonable for the calculation of population reduction rate achieved at different levels.

As indicated from the table, from upper levels of the tree to lower levels, the classification accuracy steadily decreases and the population reduction rate (PR Rate in the table) steadily increases. In 25dB case, our six-level decision tree is able to achieve a 97.06% classification accuracy and a 94.33%

TABLE I
PARAMETERS USED TO CONSTRUCT SIX-LEVEL HIERARCHICAL DECISION TREE AND PERFORMANCE ACHIEVED (25DB SNR)

Level	Feature	λ	No. of Clusters	Outlier Percentage	PR Rate(%)	Accuracy(%)
1	Pitch	0.8	16	/	51.24	99.03
2	PAR of the Positive Pulse	1.1	32	7.5%	76.61	98.50
3	Skewness of the Positive Pulse	0.55	16	2.5%	86.33	97.90
4	PAR of the Negative Pulse	0.8	16	2.5%	90.93	97.45
5	Skewness of the Negative Pulse	0.85	8	7.5%	93.01	97.27
6	Width of the Positive Pulse	0.7	16	2.5%	94.50	97.06

TABLE II
PERFORMANCE COMPARISON OF FUZZY-CLUSTERING-BASED DECISION TREE AND HARD-CLUSTERING-BASED DECISION TREE (30DB SNR)

Level	Hard-Clustering		Fuzzy-Clustering	
	PR Rate(%)	Acc.(%)	PR Rate(%)	Acc.(%)
1	85.80	96.93	89.24	99.40
2	92.29	90.51	94.95	98.74
3	94.36	82.79	97.13	98.16
4	97.50	80.37	98.10	98.06
5	98.61	67.20	98.53	97.35
6	99.20	61.00	98.83	97.16

TABLE III
PARAMETERS USED IN CALCULATIONS OF MFCC AND GMM

Window Type	Hamming
Window Length	0.0232s
Frame Rate	100 Frames/s
NFFT	256
No. of Filter Banks	31
Lowest/Highest Freq. of Filter Bank	200Hz/3500Hz
Dim. of MFCC	26
Dither	yes
Cov. Matrix of GMM	Nodal&Diagonal
Min. VAR Allowed in GMM	0.01

population reduction rate at the bottom level (i.e., each leaf node contains 216 speakers on average). The performance looks quite impressive.

To validate that the fuzzy-clustering is essential for the construction of our decision tree and it outperforms the conventional hard clustering. We compare the classification performance of the fuzzy-clustering-based decision tree and the hard-clustering-based decision tree in the AWGN scenario with 30dB SNR. We use the same six features shown in Table I to construct the two trees. The comparison is summarized in Table II. The PR rate and the accuracy are calculated as same as described above for Table I. From the table, we know that the fuzzy-clustering-based decision tree can achieve a much higher classification accuracy (Acc. in the table) than the hard-clustering-based decision tree while the population reduction achieved by the two trees are pretty much the same. For the fuzzy-clustering-based decision tree, at the bottom level, the classification accuracy is 97.16% and the population reduction rate is 98.83% (i.e., each leaf node contains 45 speakers on average). The hard-clustering-based decision tree is not applicable since a 61% classification accuracy is not acceptable.

B. Comparison with MFCC+GMM+UBM and MFCC+GMM

In this section, the performance of our fuzzy-clustering-based hierarchical decision tree approach are compared with the MFCC+GMM+UBM approach and the MFCC+GMM approach for large population speaker identification. The experiments are conducted in AWGN scenario with different SNRs and are tested with different amount of testing samples. Section V-B1 introduces the experimental settings and Section V-B2 shows the comparison in aspects of correct identification rate and computational complexity.

1) Experimental Settings for Performance Comparison: The experiments are carried out on the same dataset which has been described in Section V-A1. For the MFCC+GMM+UBM approach, we estimate the UBM with 2048 GMM mixtures by using one-hour speech of 120 male speakers (each has one 30s training sample) and one-hour speech of 120 female speakers (each has one 30s training sample). The speakers used to estimate the UBM are not used for the evaluation. In the adaptation, each speaker model is derived by adapting the parameters of UBM using the speaker's 30s training samples. We use a single adaptation coefficient for all parameters with a relevance factor of 16. In the testing phase, we use the dot-scoring technique [25] to score a testing sample against all speaker models and find the best match. For the MFCC+GMM approach and the MFCC+GMM approach invoked at leaf nodes of our hierarchical decision tree, we adopt 32-mixture GMM and for each speaker, a 30s training sample is used to estimate the 32-mixture GMM. Table III shows all parameters adopted for MFCC calculation and GMM classification in the experiments. The parameter applies to all three approaches. In our decision tree approach, the decision tree is constructed in the same way as shown in Section V-A by using the same dataset.

In VoIP, the typical SNR is at least 25dB [26]. In order to make our work realistic and applicable, our experiments are mainly carried out in the AWGN scenario with 25dB SNR and 30dB SNR. However, to validate the performance of our approach in low SNR cases, experiments are also conducted for 15dB SNR and 20dB SNR. Additionally, experiments are

carried out with shorter testing samples (3s and 10s).

2) Comparison of Accuracy and Complexity: The performance comparison in correct identification rate and computational complexity is summarized in Table IV, Table V and Table VI. The correct identification rate (CIR in the tables) is measured by the percentage of correct identification averaged across all 3805 speakers in our dataset. The computational complexity is measured by the average execution time of testing a speaker. Table IV shows the performance comparison for 25dB and 30dB SNRs when the testing sample is 30s long. As indicated from the table, we know that our fuzzy-clustering-based hierarchical decision tree approach can achieve higher correct identification rate than both MFCC+GMM+UBM approach and MFCC+GMM approach. The MFCC+GMM approach is the worst of the three approaches. For 30dB and 25dB, compared with the MFCC+GMM+UBM approach, our approach can achieve 8% and 2% increase in correct identification rate, respectively. The performance comparison is shown in Table V for 15dB and 20dB SNRs when the testing sample is 30s long. Although the performance of all approaches in low SNR cases are not satisfactory, the table indicates that our approach can achieve much higher correct identification rate than the other two GMM approaches. For 15dB and 20dB, compared with the MFCC+GMM+UBM approach, our approach can achieve 17% and 21% increase in correct identification rate, respectively. Table VI shows the performance comparison for 30dB SNR when the testing sample is 3s long and 10s long. As shown in the table, shorter testing length leads to performance reduction of all three approaches. Our approach is still the best among the three approaches. Compared with the MFCC+GMM+UBM approach, our approach can achieve about 3.5% increase in correct identification rate in both cases of 3s and 10s testing length. The great performance improvement brought by our approach comes from the significant population reduction with only a little loss of decision-tree-based classification accuracy under AWGN conditions.

In addition to the accuracy, our approach also has the big advantage in computational complexity. For speaker identification problem especially large population case, both MFCC+GMM approach and MFCC+GMM+UBM approach are very expensive. Each MFCC feature vector needs to be scored against all GMM components of all speaker models. Even the dot-scoring technique helps MFCC+GMM+UBM approach achieve lower computational complexity, there is still a large amount of likelihood computations against a great number of speaker models. Compared with the two other approaches, our approach has a much lower computational complexity because MFCC+GMM is only applied to speaker groups at leaf nodes with much smaller population size and the number of GMM mixtures is not large. Thus, the total amount of likelihood computations is greatly reduced. Our hierarchical decision tree has some additional computation including the feature extraction and the decision-tree-based classification. However, both the feature extraction and the threshold-based classification require only a very small amount of computation. Therefore, the additional computation is negligible, compared with GMM likelihood computation. Table IV, Table V and Table VI compare the average execution time of finishing testing one speaker for three approaches. As shown in the tables, although the MFCC+GMM+UBM approach is much faster than the MFCC+GMM approach, our approach required much less time than the MFCC+GMM+UBM approach. Thus, it's another good benefit of our approach. Notice that some approaches were proposed to reduce the computational complexity of the MFCC+GMM+UBM approach with only a very slight degradation of the identification performance [3], [6], [8]. Here, we do not compare the complexity of our approach with those approaches. However, we can see that the execution time of our approach to finish testing a speaker is less than or approximates the length of the testing speech, it is promising that our approach can be implemented fast enough for the real-time applications.

As a conclusion, the above comparison in both correct identification rate and computational complexity shows the superiority of our fuzzy-clustering-based hierarchical decision tree approach for large population speaker identification in AWGN scenarios.

TABLE IV
COMPARISON FOR 25DB AND 30DB SNRS (30S TESTING SAMPLE)

Approach	Avg. time per speaker(s)		CIR(%)	
	25dB	30dB	25dB	30dB
GMM	373	371	51.93	71.56
GMM+UBM	74	73	65.88	80.93
Decision Tree	30	12	67.91	88.8

TABLE V
COMPARISON FOR 15DB AND 20DB SNRS (30S TESTING SAMPLE)

Approach	Avg. time per speaker(s)		CIR(%)	
	15dB	20dB	15dB	20dB
GMM	559	520	6	20.5
GMM+UBM	103	97	5.7	33.9
Decision Tree	13	13	22.1	54.91

VI. CONCLUSIONS

As the major technique for speaker identification, approaches based on MFCC and GMM can achieve superior performance for small population identification under low-noise conditions. However, for large population identification

TABLE VI
COMPARISON FOR 3S AND 10S TESTING SAMPLE (30DB SNR)

Approach	Avg. time per speaker(s)		CIR(%)	
	3s	10s	3s	10s
GMM	75	150	53.98	66.31
GMM+UBM	51	64	56.33	75.16
Decision Tree	3	9	59.78	78.65

under noisy conditions, the performance of approaches based on MFCC and GMM suffers from severe degradation. As the population increases, the accuracy will steadily decrease and the computational complexity will proportionally increase. To improve the performance of large population speaker identification under noisy conditions, we proposed the fuzzy-clustering-based hierarchical decision tree approach. Our approach aims at using a hierarchical decision tree to partition the large population of all registered speakers into subgroups of very small population size and determining the speaker group at the leaf node to which a speaker under test belongs. Since we only use those speech features that are independent from MFCC to do speaker clustering for population partitioning, the probability of having speakers with similar MFCC is greatly reduced in speaker groups at the leaf nodes. We only apply the MFCC+GMM identification approach to the selected speaker group at the leaf node which has a small population size and hence MFCC+GMM can perform well for speaker identification with a much lower computational complexity. To achieve a low error probability of decision-tree-based classification, we proposed to adopt the fuzzy clustering rather than the conventional hard clustering in constructing the decision tree. Specifically, at each level of the tree, a speaker may belong to multiple speaker groups or nodes. Replicas of a speaker in multiple groups/nodes can greatly increase the probability of classifying the speaker (if under test) into a correct group/node in the process of decision-tree-based classification. Moreover, we developed a total of six features (including pitch and five vocal source characteristics) and constructed a six-level tree, accordingly. Experimental results have shown the excellent performance of our approach for large population identification under AWGN conditions. It is promising that our approach is applied in real-time applications of large population speaker identification under noisy conditions.

To further validate the superiority of our decision tree approach, more experiments should be conducted to test our approach on datasets with larger population in different scenarios of additive noise such as interfering speakers' voice, background music, etc. In order to accommodate larger population, more useful speech features need to be derived and more levels need to be added into the existing tree for more population reduction. Moreover, as we all know, in automatic speech recognition, it is a common practice to automatically determine the order of the features in decision tree (e.g., decision tree for context dependent acoustic modeling). For our decision tree, a feasible way to achieve the same goal, though sub-optimal, may be that we always select a feature and add it to an existing tree if a new decision tree constructed by adding one more level with this feature can achieve better classification performance than using other available features. This can be another future work for our decision tree approach.

REFERENCES

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 23–61, 2011.

[2] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.

[3] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[4] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435–1447, 2007.

[6] Z. Xiong, T. Zheng, Z. Song, F. Soong, and W. Wu, "A tree-based kernel selection approach to efficient gaussian mixture model–universal background model based speaker identification," *Speech communication*, vol. 48, no. 10, pp. 1273–1282, 2006.

[7] A. Sarkar, S. Rath, and S. Umesh, "Fast approach to speaker identification for large population using mllr and sufficient statistics," in *Communications (NCC), 2010 National Conference on*. IEEE, 2010, pp. 1–5.

[8] B. Pellom and J. Hansen, "An efficient scoring algorithm for gaussian mixture model based speaker identification," *Signal Processing Letters, IEEE*, vol. 5, no. 11, pp. 281–284, 1998.

[9] D. Reynolds, "Large population speaker identification using clean and telephone speech," *Signal Processing Letters, IEEE*, vol. 2, no. 3, pp. 46–48, 1995.

[10] V. Apsingekar and P. De Leon, "Speaker model clustering for efficient speaker identification in large population applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 848–853, 2009.

[11] U. Chaudhari, J. Navrratil, G. Ramaswamy, and S. Maes, "Very large population text-independent speaker identification using transformation enhanced multi-grained models," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 461–464.

[12] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[13] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," *Security and Communication Networks*, 2011.

[14] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 6, pp. 1097–1111, 2008.

[15] H. Ezzaidi, J. Rouat, and D. O'Shaughnessy, "Towards combining pitch and mfcc for speaker identification systems," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[16] N. Wang, P. Ching, N. Zheng, and T. Lee, "Robust speaker recognition using denoised vocal source and vocal tract features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 196–205, 2011.

[17] D. Hosseinzadeh and S. Krishnan, "Combining vocal source and mfcc features for enhanced speaker recognition performance using gmms," in *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*. IEEE, 2007, pp. 365–368.

[18] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker characterization and recognition-speaker identification and verification by combining mfcc and phase information," *IEEE Transactions on Audio Speech and LanguageProcessing*, vol. 20, no. 4, p. 1085, 2012.

[19] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2072–2084, 2007.

[20] X. Huang *et al.*, *Spoken language processing*. Prentice Hall PTR New Jersey, 2001.

[21] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[22] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, p. 1917, 2002.

[23] C. Wang, "Prosodic modeling for improved speech recognition and understanding," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[24] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition. i," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 29, no. 6, pp. 778–785, 1999.

[25] M. Diez, M. Penagarikano, A. Varona, L. Rodriguez-Fuentes, and G. Bordel, "On the use of dot scoring for speaker diarization," *Pattern Recognition and Image Analysis*, pp. 612–619, 2011.
[26] L. Narváez, J. Pérez, C. García, and V. Chi, "Designing 802.11 wlans for voip and data," *IJCSNS*, vol. 7, no. 7, p. 248, 2007.

Yakun Hu received B.E. in Electronic and Information Engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2005, M.E. in Communication and Information Systems from Southeast University, Nanjing, China, in 2008, and Ph.D. in Electrical and Computer Engineering from University of Florida, Gainesville, FL, in 2012.

His research interests include speaker recognition, signal processing, machine learning and wireless communications.

Dapeng Oliver Wu (S'98–M'04–SM06–F'13) received B.E. in Electrical Engineering from Huazhong University of Science and Technology, Wuhan, China, in 1990, M.E. in Electrical Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 1997, and Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, PA, in 2003.

Since 2003, he has been on the faculty of Electrical and Computer Engineering Department at University of Florida, Gainesville, FL, where he is a professor; previously, he was an assistant professor from 2003 to 2008, and an associate professor from 2008 to 2011. His research interests are in the areas of networking, communications, signal processing, computer vision, and machine learning. He received University of Florida Research Foundation Professorship Award in 2009, AFOSR Young Investigator Program (YIP) Award in 2009, ONR Young Investigator Program (YIP) Award in 2008, NSF CAREER award in 2007, the IEEE Circuits and Systems for Video Technology (CSVT) Transactions Best Paper Award for Year 2001, and the Best Paper Awards in IEEE GLOBECOM 2011 and International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks (QShine) 2006.

Currently, he serves as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology, Journal of Visual Communication and Image Representation, and International Journal of Ad Hoc and Ubiquitous Computing. He was the founding Editor-in-Chief of Journal of Advances in Multimedia between 2006 and 2008, and an Associate Editor for IEEE Transactions on Wireless Communications and IEEE Transactions on Vehicular Technology between 2004 and 2007. He is also a guest-editor for IEEE Journal on Selected Areas in Communications (JSAC), Special Issue on Cross-layer Optimized Wireless Multimedia Communications. He has served as Technical Program Committee (TPC) Chair for IEEE INFOCOM 2012, and TPC chair for IEEE International Conference on Communications (ICC 2008), Signal Processing for Communications Symposium, and as a member of executive committee and/or technical program committee of over 80 conferences. He has served as Chair for the Award Committee, and Chair of Mobile and wireless multimedia Interest Group (MobIG), Technical Committee on Multimedia Communications, IEEE Communications Society. He is a member of Multimedia Signal Processing Technical Committee, IEEE Signal Processing Society. He is an IEEE Fellow.

Antonio Nucci is the Chief Technology at Narus. His research interest lies in traffic measurement and engineering, network security and forensics. In his career his published more than 90 ACM/IEEE papers, 31 patents awarded with USPTO and co-authored a definitive textbook titled, "Design, Measurement and Management of Large-Scale IP Networks: Bridging the gap between Theory and Practice" published by Cambridge University Press. He serves as a Technical Advisor of several venture capital firm in Silicon Valley including Panorama Capital LLC (spin-off of JP Morgan) advising on emerging technologies and trends.

Dr. Antonio Nucci obtained his Ph.D. and Master's Degrees in Electrical Engineering from Politecnico di Torino, Italy.