# Caller Identification for the SCANMail Voicemail Browser

*Aaron Rosenberg*[1], *Julia Hirschberg*[1], *Michiel Bacchiani*[1],
*S Parthasarathy*[1], *Philip Isenhour*[2], *Larry Stead*[1]

AT&T Labs-Research[1], Virginia Tech[2]
aer@research.att.com

## Abstract

SCANMail is a prototype system developed at AT&T Labs for the purpose of providing useful tools for managing and searching through voicemail messages. Content is extracted from voicemail messages using various speech and text processing tools. One such content category is the identity of the message caller. This paper describes CallerID, the server tool attached to SCANMail for the purpose of providing caller labels for voicemail messages. CallerID makes use of text independent speaker recognition techniques. Two kinds of requests are handled by the CallerID server. A request triggered by the arrival of a new voicemail message results in the processing of the message to score it against the models of callers assigned to the user (recipient) in order to propose the identity of the caller. A second request is initiated by a user who provides a caller label for a message he/she has reviewed. CallerID processes the message and uses it to train or adapt a speaker model for the caller whose label is provided. The paper describes in detail the CallerID functions and provides some results of performance evaluations of the caller identification capability.

## 1. Introduction

Ever increasing amounts of data bandwidth and storage have provided us with access to huge amounts of multimedia information. The capability for managing, searching, and browsing through this information, however, is mixed. For text information, sophisticated search engines and other information retrieval tools offer convenient and reliable facilities for finding desired sources of information and organizing them for retrieval. Analogous facilities for audio and video information are severely deficient at present. As an example, good text-based tools exist for browsing through and organizing email messages. However, comparable tools for voicemail messages have been heretofore largely nonexistent. SCANMail is a prototype system developed at AT&T Labs for the purpose of providing useful tools for managing and searching through voicemail messages. It employs automatic speech recognition (ASR) to provide text transcriptions, information retrieval on the transcription to provide a weighted set of search terms, information extraction to obtain key information such as telephone numbers from transcription, as well as automatic speaker recognition to carry out caller identification (CallerID) by processing the speech data. It also employs a set of human computer interaction tools via a graphical user interface (GUI) which enable a user to exercise features of the system. A detailed description of the system can be found in a companion paper at this conference [1].

This paper focuses on the CallerID tool for SCANMail. Automatically supplying caller labels to voicemail messages can facilitate user access to the desired information content in voicemail messages with such queries as "Play the latest message from $[caller]$" or "Find a message about $[topic]$ sent by $[caller]$". The CallerID server proposes caller labels by matching processed messages against existing caller models assigned to a user. Caller models are created from training data labeled by users in the course of reviewing their messages.

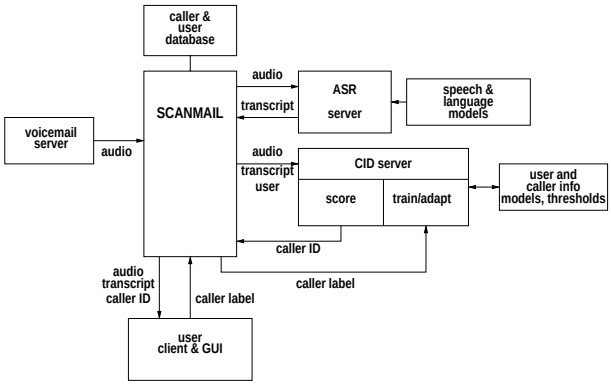


Figure 1: *Functional diagram of CallerID processing in SCANMail.*

There are other possible sources of information for identifying a caller. For many types of calls, for example, calls within a local PBX, the originating telephone number may often be available in the voicemail header. It may also be possible to extract the caller's name and/or telephone number from the message, when provided, using ASR. These sources of information could be combined with the speaker recognition hypothesis to propose a caller label.

## 2. CallerID functions in SCANMail

A functional diagram of CallerID processing in the SCANMail system is shown in Figure 1. A more detailed view of the CallerID processing is shown in Figure 2. Voicemail messages in the SCANMail prototype are retrieved from a commercial voicemail system, Audix$^{TM}$, provided by Avaya Corporation, which is attached to the local PBX. Two types of processing requests are handled by the CallerID server. For the first type, SCANMail polls the voicemail server for the arrival of a new message for one of the users. When a new message arrives, a request is made to the ASR server to process the message audio signal. The ASR server returns a time-aligned transcription of the message. SCANMail then requests processing by the CallerID server. The input to CallerID consists of the message audio signal and transcription and the identity of the message recipient or SCANMail user. The transcription is used only to

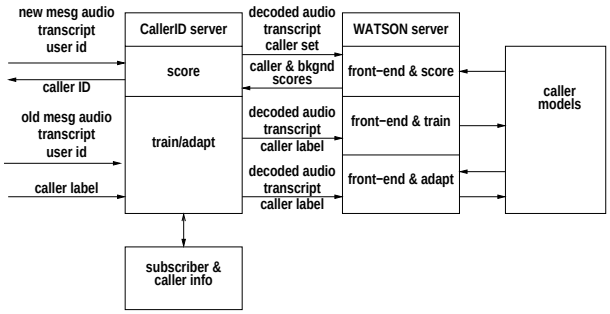


Figure 2: *The CallerID functions in detail.*

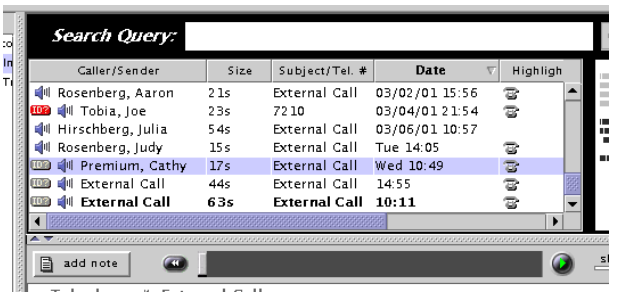


Figure 3: *Portion of the SCANMail user GUI which lists messages. An explanation of CallerID features is provided in the text.*

segment the audio signal into speech and nonspeech segments. The CallerID server compares the processed speech segments of the audio signal with the model of each caller in the user's caller set. These callers are selected when the user supplies a caller label to messages he/she has reviewed, as explained below. A matching score is obtained for each such comparison. The best matching score is compared with a caller dependent rejection threshold. If the matching score exceeds the threshold, the CallerID server reports the corresponding caller label. Otherwise, the CallerID report is "unknown caller". This information is relayed back via the SCANMail system the user's client and displayed on his/her GUI.

The second type of request for CallerID processing originates with the user. In the course of reviewing (playing back) his/her messages, the user has the capability, via the GUI, of supplying a caller label to a message previously labeled as "unknown" by CallerID, or to confirm or correct a label supplied by CallerID or extracted from the message header. An example of the portion of the GUI which lists messages is shown in Figure 3. The last message in this list is printed in bold indicating that it has not been reviewed (played). The "caller/sender" field shows "External Call" which indicates that CallerID has labeled the message as "unknown" [1] When the user clicks on the "ID?" icon on the left, he/she will be provided with drop-down panels and allowed to identify the caller. This can be done by selecting a caller from a list of callers already assigned to the user or by typing identifying information in specified fields. Some other messages in the list with the "ID?" icon have a caller label either provided by CallerID or associated with the originating telephone number. The user can confirm or correct these labels, again, by clicking on the icon. The messages without the icon have already been confirmed or corrected. Each such user-supplied caller label generates a request to the CallerID server to use the labeled message as training data to construct or adapt a speaker model for the associated caller. Caller models and training data are shared among users.

As indicated in Figure 2, most of the CallerID scoring, training and adaptation processing take place in Watson, the AT&T Speech Recognition Engine[2], acting as a server. For scoring, the CallerID server decodes the low complexity CELP (LC-CELP) Audix$^{TM}$ message audio signal (retrieved from the voicemail store) and passes it on to Watson for front-end analysis and the calculation of log likelihood scores for both caller and background models, as described in more detail in Section 3. The scores are passed back to the CallerID server for normalization and an identification decision. Similarly, for training and adaptation, front-end analysis and training and adaptation processing take place in Watson, while the CallerID server manages and stores the training or adaptation data.

[1] In the current implementation, the CallerID hypothesis is shown only if the PBX cannot supply a unique name to the originating telephone. This occurs for all external calls and, internally, when the originating telephone is not assigned to a person.

## 3. Speaker recognition processing

Speaker recognition is carried out using text independent techniques in which callers are represented by 64-component Gaussian mixture models (GMM's) [3]. The front-end processing applies a mel-spaced filter bank cepstral analysis to decoded LC-CELP Audix$^{TM}$ messages. The ASR supplied time-aligned transcription segments the audio into speech and nonspeech segments. Descriptions of the model construction and front-end processing can be found in an earlier paper [4]. Likelihood ratio calculations are used to obtain matching scores for caller identification. Log likelihood scores are computed for each feature vector frame of a processed message with respect both to a target caller model and to speaker background models. The log likelihood scores are averaged over all the speech frames. The matching score is the average normalized score or log likelihood ratio score for a caller $T$ and is obtained as

$$SN(X|\lambda_T) = S(X|\lambda_T;\lambda_{B_1},\lambda_{B_2},\ldots,\lambda_{B_K}) = S(X|\lambda_T) - \max_k S(X|\lambda_{B_k}) \quad (1)$$

where $S(X|\lambda)$ is the average log likelihood score for model $\lambda$ over all $N$ frames $X = \{x_1, x_2, \ldots, x_N\}$ of the speech portions of the signal. $B_k$ refers to the $k$-th speaker background model. The matching scores for all callers assigned to the user are sorted to find the best matching score. A decision is made whether or not to label the message with the caller associated with the best matching score. In addition to the best matching score, if the number of callers assigned to a user is sufficiently large, the difference between the best and next best matching score is also used to make the decision. The scores are compared with caller dependent thresholds for both the best matching score and difference between the best and next best matching score. If either threshold test succeeds, the caller label is supplied; otherwise the message is labeled as "unknown". Thresholds are allowed to adapt from trial to trial using an empirical maximum *a posteriori* (MAP) formulation based on the current threshold and the score history. Thresholds are both caller- and user-dependent.

At the outset, a new user has no assigned callers. In this state, by default, all messages are labeled as "unknown" by the CallerID server. As described in the previous section, the user

has the option of supplying a caller label to any message labeled as "unknown" via the user's GUI. When the accumulated duration of user-labeled messages for a caller exceeds 60 secs, an initial caller model is created. In the current implementation, when an additional 20 secs of messages is accumulated, a new model is created using all the training data accumulated so far. This process continues until 180 secs of messages have been accumulated. Thereafter, a user-labeled message can be used to adapt the current model. The feature vectors in the message are used to update GMM component means and weights using a MAP technique described in an earlier paper [5].

## 4. Experimental performance evaluations

The nature of the application presents distinct challenges for automatic speaker recognition. First, the application is an open-set identification problem. A typical user will have a relatively small set of callers who call regularly and leave messages for whom caller identification would be useful. We refer to this set as the user's "ingroup". The user can also expect messages from a larger set of callers who do not leave messages regularly for whom caller identification may not be as important. We refer to this set as the user's "outgroup". It is important that "outgroup" callers, as "imposters", not be accepted as "ingroup" callers, the "true customers". Although it is not clear what the *a priori* probabilities of occurrence are for ingroup and outgroup calls are in practice, the probability of outgroup acceptance inevitably increases as the ingroup becomes larger.

A second challenge is the expected variety of recording and transmission conditions for voicemail messages. Ordinary telephone handsets, both electret and carbon button microphones, speakerphones, and cellular phones and various recording environments can be expected to produce a wide variety of recording conditions. Models and scores need to be robust with respect to these expected variations.

Although most voicemail messages can be expected to include only one talker, some, including some forwarded messages, may have more than one talker. We do not deal with this issue at present, excluding such messages from our evaluations and instructing users not to label such messages in our prototype system.

Three kinds of possible identification errors can be obtained. An outgroup caller can be identified as an ingroup caller, an ingroup caller can be accepted as another ingroup caller, and an ingroup caller can be labeled as unknown. These are referred to as outgroup acceptance, ingroup confusion, and ingroup rejection, respectively.

Two performance evaluations, one large-scale and a second small-scale, have been carried out. In the large-scale evaluation, an experimental database is extracted from a corpus of approximately 10,000 voicemail messages collected from the mailboxes of approximately 140 AT&T Labs employees over a 3-month period. These messages were recorded and digitized at an 8kHz sampling rate as 8-bit mulaw samples on a voicemail system at AT&T Labs which is no longer used. The messages were transmitted over a representative variety of telephones including ordinary telephone handsets, speakerphones, and cellular phones. The messages are manually labeled, including information about the caller. The name of the caller, if provided in the message, is included as a label, as well as such information as gender, age (child/adult), foreign language, speech pathology, etc. The average message duration is about 40 secs and the median, 25 secs. Messages used for the evaluation are those for which it can be determined that the caller is uniquely labeled. Three groups of messages are selected. The first group consists of 973 messages from 20 distinct callers, 11 female and 9 male, designated as ingroup. Each ingroup caller has at least 20 messages and a total message duration of at least 10 minutes. The first 6 minutes of each ingroup caller's messages is reserved for training and the balance is used for testing. A set of 220 messages, each from a distinct caller not included in the ingroup set, is designated outgroup. 130 outgroup callers are male and 90 are female. A third distinct set of 138 messages from each of 138 callers, approximately half male and half female is used for background model training. Each of these messages is truncated to 15 secs. This set is further divided into a set of 92 messages from ordinary telephone handsets and the remainder from other types of phones, to construct background models representing these two conditions. A number of experimental variables have been examined in this evaluation including the number and kind of background models, the number of ingroup callers, the front-end processing (LPC-derived vs. mel-spaced filter bank cepstral coefficients), the length of test messages, speaker independent and speaker dependent decision thresholds, and the training and adaptation state of caller models. The experimental results are described in detail in a previous paper [4].

A representative set of results is shown in the first row of Table 1. Here, two background models are used, one, obtained from ordinary handset training data, the other from non-handset training data, the front end is based on mel-spaced filter bank cepstral coefficients, the decision thresholds are speaker dependent, the models use 4 mins of training data and are adapted with data from rejected test utterances. Threshold parameters have been set to maintain outgroup acceptance at a relatively low level of 2.7% at the expense of a relatively high level, 11.0%, of ingroup rejection. Ingroup confusion is low, at 1.2%.

| experiment | ingroup rejection | ingroup confusion | outgroup acceptance |
|---|---|---|---|
| 20-caller | 11.0 | 1.2 | 2.7 |
| 14-caller | 8.3 | 0.0 | 0.9 |

Table 1: *Average caller identification error rates (%) obtained in two performance evaluations. The 20-caller evaluation uses 4-min. caller models adapted with rejected messages. The 14-caller evaluation uses 1-min., unadapted, caller models. Decision thresholds are caller dependent and allowed to adapt from trial to trial within limits.*

In the small-scale, evaluation, 14 male speakers recorded at least 5 messages on the current Audix$^{TM}$ voicemail system. The first two messages, at least 30 secs each, are reserved for training, and the remainder, at least 10 secs each, are reserved for testing. The primary goal of this small-scale evaluation is to validate performance using the LC-CELP encoded messages of this voicemail system. Models are trained on 1 minute of data. There is no model updating. There are 48 ingroup test messages, 3 or 4 from each caller. The outgroup messages are the same set of 220 messages used for the large-scale evaluation. Results are shown in the second row of Table 1. Performance is significantly better than that obtained for the 20-caller evaluation, with 0.9% outgroup acceptance and 8.3% ingroup rejection. There is no ingroup confusion. Although better performance can be expected with a smaller ingroup, it seems likely that much of the improvement can be attributed to the fact that all the ingroup test messages originate from the same type of handset within the local PBX at AT&T Labs. Note that better performance is obtained here even though the 14-caller models

are created with 1 min of training data, compared with 4 mins. of training data for the 20-caller models, and are unadapted.

## 5. Discussion

SCANMail is a system which extracts useful content from voicemail messages to enable users to browse and search through their messages with many of the same capabilities available to email users. One such content category is the identity of the message sender. CallerID is a SCANMail component, employing automatic speaker recognition techniques to extract speaker characteristics embedded in a voicemail message speech signal, which hypothesizes caller labels for the messages. It represents a novel application for automatic speaker recognition which, until recently, has mainly been proposed for security applications, using voice characteristics as a kind of personal identity information to control access to secure premises, information, transactions, etc. One troublesome issue in automatic speaker recognition is the acquisition of labeled training data. In security applications, training data acquisition must be carried out in a carefully supervised manner and, for the convenience of the customer, with a small amount of data collected in as short amount of time as practicable. These constraints generally necessitate a single enrollment session. This has the distinct disadvantage of providing just one, possibly not well representative, sample of training data.

In contrast, in the voicemail application, training data is labeled by the user. For each caller, the data comprise a set of distinct messages and is thus likely to be more representative of the expected variability. Moreover, if an unrepresentative message from a caller is rejected and subsequently labeled by the user, that data will be folded in the caller model making it more representative. It will then be more likely that a subsequent similar message is identified correctly.

As previously discussed, the voicemail application also creates some significant challenges. First, the application is open-set identification. It is not clear what the largest practicable ingroup size might be. Each additional caller in the ingroup increases the possibility of outgroup acceptance. Finally, the application is, of course, telephone based, with all the recording and transmission variations that that implies. Although our processing includes many features that help compensate for the variability [4], error rates are somewhat high. The differences in performance between the large-scale evaluation, which contains most of the expected sources of variability, and the small-scale evaluation, where channel and recording variability is significantly restricted, illustrates the problem.

Future activities for voicemail speaker recognition include improvements in the CallerID server features, such as the ability to repair caller models in response to user correction of labeling errors, and some basic improvements in speaker recognition techniques for this application. In addition, an upcoming 20-subject user acceptability test, which includes logging all user behavior automatically and written and oral surveys, will provide information on the utility and convenience of SCANMail features, including CallerID. It will provide some data on the size of user ingroups and the frequency of ingroup and outgroup calls as well as overall CallerID performance.

## 6. References

[1] Julia Hirschberg, Michiel Bacchiani, Don Hindel, Phil Isenhour, Aaron Rosenberg, Litza Stark, Larry Stead, Steve Whittaker, and Gary Zamchick, SCANMail: Browsing and searching speech data by content, *Proc. Eurospeech 2001*, Aalborg, September, 2001.

[2] R.D. Sharp, E. Bocchieri, C. Castillo, S. Parthasarathy, C. Rath, M. Riley, and J. Rowland, "The Watson Speech Recognition Engine," *Proc. ICASSP 97, 1997 Intl. Conf. on Acoustics, Speech, and Signal Processing*, 4065-4068, Munich, April, 1997.

[3] D. A. Reynolds and R. C. Rose, Robust text-independent speaker identification using Gaussian mixture models, *IEEE Trans. on Speech and Audio Processing*, **3**, 72-83, 1995.

[4] Aaron E. Rosenberg, S. Parthasarathy, Julia Hirschberg, and Stephen Whittaker, Foldering voicemail messages by caller using text independent speaker recognition," *Proc. ICSLP 2000, Sixth Intl. Conf. on Spoken Language Processing*, **II**, 474-477, Beijing, October, 2000.

[5] A.E. Rosenberg and F.K. Soong, Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes, *Computer Speech and Language*, **2**, 143-157, 1987.