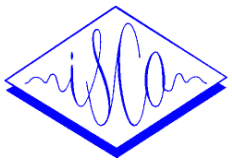

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

FOLDERING VOICEMAIL MESSAGES BY CALLER USING TEXT INDEPENDENT SPEAKER RECOGNITION

Aaron E. Rosenberg *S. Parthasarathy* *Julia Hirschberg* *Stephen Whittaker*

AT&T Labs-Research
Florham Park, NJ 07932 USA

ABSTRACT

The ability to automatically scan voicemail messages for content and caller identity cues would be a useful service. This paper describes a system which automatically files voicemail messages into caller folders using text independent speaker recognition techniques. Callers are represented by Gaussian mixture models (GMM's). The speech for an incoming message is processed and scored against caller models created for a subscriber. A message whose matching score exceeds a threshold is filed in the matching caller folder; otherwise it is tagged as "unknown". The subscriber has the ability to listen to an "unknown" message and file it in the proper folder, if it exists, or create a new folder, if it does not. Such subscriber labelled messages are used to train and adapt caller models. The system has been evaluated on a database of voicemail messages collected at AT&T Labs. A set of 20 callers from this database is designated as "ingroup". Each of these callers has recorded at least 20 messages totalling 10 or more minutes in duration. A distinct set of 220 messages, each from a different caller, are designated as "outgroup". Representative performance figures with threshold parameters set to ensure that outgroup acceptance is low compared with ingroup rejection are the following. The average ingroup message rejection rate is 11.0% and the average ingroup message confusion rate (matching the wrong caller) is 1.0%, while the average outgroup message accept rate is 2.7%.

1. INTRODUCTION

The search and navigation capabilities that allow subscribers to manage and sort email messages are not available for voicemail. Email messages come tagged with header information about the sender and the subject. Both the header and body text of such messages can be readily searched using simple key word or information retrieval techniques to locate messages about a particular topic and/or sent by a particular person. Email text filters can sort messages into appropriate folders even before they are read. We are currently engaged in research designed to provide search and navigational capabilities for voicemail messages. The speech content of each message can be analyzed to identify words or phrases as well as speakers.

This paper describes a system for identifying voicemail message callers by applying text independent speaker recognition techniques to the speech content in the messages. The speech for each incoming message is processed and compared with existing speaker models established for the subscriber. If the message does not score well enough against the existing speaker models, or if no models exist, it is tagged as "unknown". When the subscriber listens to a message tagged as unknown, he/she has the opportunity to label it and file it in the proper folder, if it exists, or create a new caller folder, if it does not. When the accumulated duration of messages stored in a folder is sufficient, an initial speaker model is created for the caller. Subsequent messages for the caller are stored in the folder if the matching score is good enough. New messages from the caller, together with old messages, may be used to retrain the speaker model. Messages from the caller that are rejected (tagged as unknown) or tagged incorrectly with another caller's label, but corrected by the subscriber after listening to the message, may be used to adapt caller models.

2. DATABASE DESCRIPTION

The experimental database is extracted from a corpus of approximately 10,000 voicemail messages collected from the voice mailboxes of approximately 140 employees at AT&T Labs over a 3-month period. The messages were transmitted from a representative variety of telephones including ordinary telephone handsets, speakerphones, and cellular phones. The recorded messages are digitized at an 8 kHz sampling rate as 8-bit mulaw samples. Each message is manually labelled, 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. A histogram showing the distribution of individual message durations is shown in Fig. 1.

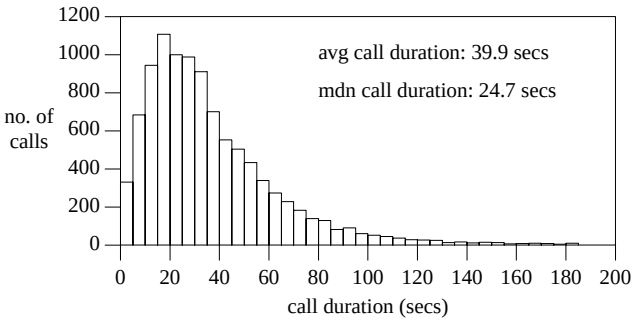

Figure 1. Histogram of individual message durations

For the purposes of the experimental evaluation, messages are selected for which the caller label contains both a first and last name to ensure that each caller label is associated with a unique caller. Also, no messages are selected in which more than one person is speaking. The experimental messages are divided into four groups. The first group, designated "ingroup", consists of 973 messages recorded by 20 callers, 11 female and 9 male adults. Each ingroup caller has at least 20 messages with a total duration of at least 10 minutes. Ingroup caller set size is an experimental variable in the evaluation. A set of 220 messages, each from a distinct caller (not included in the ingroup messages), 130 male and 90 female, are designated "outgroup". Finally, two distinct groups of messages, one drawn from 43 callers and the other from 138 callers, both approximately half male, half female, are used to create speaker background models.

3. DETAILS OF OPERATION

3.1. Front-end processing

Two types of front-end processing are used in the experimental evaluations. The first is based on linear predictive coding (LPC) based cepstral analysis, the second on mel spaced filter bank cepstral analysis. Each digitized message is trimmed by approximately 0.25 secs at the beginning and end to avoid recorded onset and offset clicks. For the LPC-based analysis, 10th order LPC coefficients are calculated every 10 ms over 30 ms windows. The LPC coefficients are converted to 12th order cepstral coefficients and augmented by 12th order delta cepstral coefficients calculated over 5-frame windows. Frames with energy falling below 30 dB below the peak energy over the message are eliminated. Channel normalization is carried out by means of utterance based cepstral mean subtraction.

For the mel filter bank analysis, 12th order cepstral coefficients are calculated by applying a discrete cosine transform (DCT) to the output of 24 mel scale spaced filters every 10 ms over 20 ms windows throughout the digitized message. Real time energy normalization and cepstral bias removal are applied with a 300 ms look ahead window. The cepstral coefficients are augmented by 12th order delta-cepstral plus delta-delta-cepstral coefficients plus energy, delta-energy and delta-delta-energy.

3.2. Training and models

Callers are represented by 64-component Gaussian mixture models (GMM's) iteratively trained using the Expectation-Maximization (EM) algorithm, initialized with the segmental K-Means algorithm [?]. Each caller's messages, arranged sequentially, are partitioned into two groups. Approximately the first six minutes of messages are designated for training models. The balance of the messages (at least 4 minutes cumulative duration) are used for testing. Several schemes for training caller models have been examined. The scheme reported here is the following. An initial model is created for a caller when at least 60 secs of messages have been accumulated. (The longest message may be truncated so that the total accumulated duration does not exceed 90 secs.) These are referred to as the first stage of training messages. When an additional 60 secs of messages are accumulated, a new model is created containing both the first and new stage of messages. This process continues until 4 stages of messages are accumulated. The fourth stage model is referred to as "mature". This sequential process of model building has been found to help create adequately representative caller models.

In addition to caller models, speaker background models are created. These are constructed from the "background" message lists mentioned in Section 2. Single background models are constructed from the 43-caller set and the 138-caller set. Each caller's messages are truncated to 15 secs for a total of 645 secs for the first set, referred to as b43, and 2070 secs for the second set, referred to b138. In addition, a set of 2 background models is created from the 138-caller set. Each message in this set is listened to to judge whether it originates from an ordinary telephone handset (with electret or carbon button microphone) or from a speakerphone or cellular phone. 92 15-sec messages, totalling 1380 secs, are used to create an ordinary handset speaker background model while 46 messages, totalling 690 secs, are used to create a non-handset speaker background model. This set is referred to b138-2.

3.3. Scoring

Scoring a test message proceeds as follow. Let $X = \{x_1, x_2, \ldots, x_N\}$ be a sequence of feature vectors representing a processed test message. Let $\lambda_{T_1}, \lambda_{T_2}, \ldots, \lambda_{T_M}$ be GMM's for each of a set of M ingroup callers and $\lambda_{B_1}, \lambda_{B_2}, \ldots, \lambda_{B_K}$ be the set of K speaker background models. (In these experiments K is either 1 or 2.) Log likelihood scores are computed for test frame x_t with respect to a model λ,

$$s(x_t|\lambda) = \log p(x_t|\lambda) \tag{1}$$

and averaged over all the processed message frames

$$S(X|\lambda) = \frac{1}{N}\sum_{t=1}^{N} s(x_t|\lambda) \tag{2}$$

The average normalized score for an ingroup caller T is obtained as

$$SN(X|\lambda_T) = S(X|\lambda_T; \lambda_{B_1}, \lambda_{B_2}, \ldots, \lambda_{B_K}) = S(X|\lambda_T) - \max_k S(X|\lambda_{B_k}) \tag{3}$$

A normalized score $SN(X|\lambda_{T_j})$ is obtained for each ingroup caller model $j = 1, 2, \ldots, M$.

3.4. Identification

The normalized scores for a message are sorted and ranked. The ingroup caller T_{r_1} associated with the best scoring caller model is tentatively identified as the caller. The best ranking score is compared with a caller dependent threshold $THN(T_{r_1})$. If the score exceeds the threshold, the identification is confirmed. Otherwise it is rejected, tagging the message as "unknown". If the number of callers, M, is greater than some preassigned value M_c (currently taken to be 10), an additional test is performed. A difference score

$$SD(X; T_{r_1}, T_{r_2}) = SN(X|\lambda_{T_{r_1}}) - SN(X|\lambda_{T_{r_2}}) \tag{4}$$

between the best and next best scores is calculated and compared with a threshold, $THD(T_{r_1})$. In this case, if either threshold test succeeds, the identification is confirmed; otherwise it is rejected. The use of difference scores is restricted to ingroup caller set sizes greater than M_c in order to assure statistical stability.

In some experiments thresholds are allowed to adapt from trial to trial according to an algorithm in which the next threshold is function of the current threshold and the current score. Initial thresholds are set empirically.

3.5. Model adaptation

Caller models can be updated when a message is rejected. This simulates the situation in which a message is reported to the subscriber as "unknown" and the subscriber subsequently files the message in the proper caller folder. In the model adaptation, the message feature vectors are used to adapt the means and mixture weights for the caller's GMM using a procedure similar to what is described in [?].

4. EXPERIMENTAL EVALUATIONS

Experimental evaluations are carried out assuming that a single subscriber has assigned caller folders to the entire set of 20 ingroup callers or to some subset of it. There are a total of 734 test messages from the ingroup callers. The number of messages per caller varies widely from 11 to 116 with an average of 36.7 and a median of 27.5. It is also assumed that the subscriber receives 220 messages from the outgroup callers for whom no folders are assigned. In a typical experiment all the subscriber's ingroup messages and all the outgroup messages are scored against all existing ingroup caller models and the background models.

The following performance statistics are calculated for each experiment. For ingroup messages, performance statistics are calculated either over the entire set of messages or over each caller's messages and then averaged over the set of callers. Closed-set error rate is the fraction of messages for which the correct caller is not the best match.

Ingroup reject rate is the fraction of messages which fail the decision threshold (whether or not the correct caller is the best match). Ingroup confusion rate is the fraction of ingroup messages for which the wrong caller is accepted. For outgroup messages, false accept rate is the fraction of outgroup messages which are accepted.

The following experimental variables are examined: the number and kind of background models, the number of ingroup callers, the front-end processing (LPC derived cepstral coefficients or mel-spaced filter bank cepstral coefficients), the length of test messages, speaker independent and speaker dependent decision thresholds, and model adaptation. Most performance statistics will be shown for so-called "mature" caller models where each caller model is trained from a total of 4 to 6 mins of messages drawn from the training message lists. For speaker independent thresholds, a fixed threshold (or set of thresholds) is found which minimizes the overall ingroup reject rate when the outgroup accept rate is 5%. In addition, performance will be described when caller models are bootstrapped from scratch.

4.1. Ingroup caller set size

First, we compare performance as a function of the size of the ingroup caller set. In this experiment, a single background model, b43, is used, the front end uses LPC-derived cepstral coefficients, and full length messages are scored. The performance is shown in Table 1 for caller set sizes 20, 10, 5, and 1. In this and subsequent tables, two error rate figures are shown in each box separated by a slash. The first represents errors averaged over messages and the second errors averaged over callers. For caller set size 10, 10 different caller sets are constructed by selecting callers evenly distributed from the entire set of 20. Each caller is included in 5 different sets. Similarly, 20 size 5 caller sets are constructed with each caller included in 5 different sets. Only best matching scores are used for caller set sizes 5 and 1. Both reject and closed-set error rates decrease as ingroup size decreases. Ingroup confusion is negligible for all ingroup sizes. Ingroup size 1 is equivalent to a verification mode and the reject rate is approximately at the verification equal-error rate.

ingroup size	20	10	5	1
reject	22.6/22.8	18.6/19.9	14.2/15.2	4.4/5.7
confusion	0.0/0.0	0.2/0.1	0.0/0.0	–
closed set	5.2/4.8	3.2/3.1	2.0/2.2	–

Table 1. Average ingroup error rates (%) as a function of group size when outgroup acceptance is set at 5%.

4.2. Front end processing, message length, background models

Ingroup error rates are shown in Table 2 for a variety of conditions. Ingroup confusions rates, again, are quite small. However, decreases in ingroup reject rates are often accompanied by small increases in ingroup confusion rates.

front end	mesg length	back model	closed set	ingroup reject	ingroup confuse
lpc cep	whole	b43	5.2/4.8	22.6/22.8	0.0/0.0
lpc cep	whole	b138	5.2/4.8	20.4/23.6	0.4/0.2
lpc cep	whole	b138-2	5.2/4.8	18.0/22.0	1.0/0.3
lpc cep	trunc	b138-2	5.9/4.7	21.0/26.1	1.2/0.6
mel cep	whole	b43	7.5/6.1	19.1/16.9	1.0/0.6
mel cep	whole	b138-2	7.5/6.1	16.4/15.2	1.5/0.8

Table 2. Average ingroup error rates (%) for various conditions when outgroup acceptance is 5%.

The first three rows compare performance for different background models with LPC-cepstrum front end processing. Using ingroup reject rates to compare performance, it can be seen that there appears to be a slight improvement for b138, the large background speaker set, compared to b43, the smaller one. There is a more distinct improvement for the 2-model large speaker set, b138-2. Recall that one model in this set represents ordinary handsets while the other model is a catch-all for speakerphones, cellular phones, etc. The scoring normalization process selects the best matching of these two background models for normalization (see Sec. 3.3.). This suggests that improved performance can be obtained by using a set of background models each of which represents the different calling conditions expected for a message. The last two rows compare background models (b43 and b138-2) for the mel cepstrum front end with similar conclusions.

Comparing now LPC cepstrum and mel cepstrum front ends, it can be seen that although for closed-set error rate, mel cepstrum performs worse then lpc cepstrum, the reverse is true for open-set error rates. The open-set improvement can be attributed to a more homogeneous distribution of scores across the population of callers for mel cepstrum processed messages. This results in more compact and better separated distributions for ingroup and outgroup scores.

Also shown in Table 2 is a performance comparison between scoring whole messages and scoring messages which are truncated to 20 secs, which is slightly less than the median duration of the messages (see Fig. 1). There is about a 15% degradation in ingroup reject rates with truncated messages.

condition	ingroup reject	ingroup confusion	outgroup accept
lpc cep	18.3/19.9	1.0/0.3	6.8
mel cep	17.0/16.8	1.6/0.6	2.3
mel cep w/updates	11.0/12.8	1.2/0.5	2.7

Table 3. Average error rates (%) with speaker dependent thresholds for mature models. Thresholds are allowed to adapt from trial to trial.

4.3. Speaker dependent thresholds

Table 3 compares performance when speaker dependent thresholds are used. Here score thresholds for all callers are initially set to the same values and allowed to adapt from message to message if the message is accepted. The threshold adaptation parameters are set so that the ingroup reject rates are approximately the same as those obtained with speaker independent thresholds. The values of updated thresholds depend on the current threshold and the score for the current message. Outgroup messages are scored after all the ingroup messages. The initial thresholds for outgroup messages are the final thresholds for ingroup messages. Outgroup messages are also allowed to update thresholds. The first two rows compare performance between lpc cepstrum and mel cepstrum front ends. At those settings it can be seen that the outgroup accept rate for mel cepstrum is less than half the rate obtained with the lpc cepstrum. Finally, the last row shows the effect of model adaptation. Models are updated using data from messages that have been rejected. Starting with the same threshold parameters used in the experiment with no model updating, it can be seen that overall ingroup rejection is reduced by some 25% or 30% with only a slight increase in outgroup accept rate.

4.4. Enrollment and adaptation

The experiments described so far show performance for test messages scored against so-called mature models (models trained on 4 to 6 minutes of training messages). It is also important to consider performance as models are trained, simulating the situation in which a subscriber, starting from enrollment, adds new caller folders to his/her list. Figure ?? shows performance for such a scenario. Error rates are shown as a function of "rounds". In each round a set of ingroup training messages followed by the entire

set of outgroup messages is scored against existing caller models. At the outset, no caller models exist. In turn, training messages from each of the first 3 ingroup callers are used to train models successively through maturity (see Sec. 3.2.) and scored against caller models as they become available. In round 2, the next 3 callers are trained, and so forth, through round 7 after which all 20 caller models are trained. In round 8, the ingroup test messages followed by outgroup messages are scored against all 20 models. Since, the number of training messages for each round is small, averaging 23, the ingroup error rates fluctuate significantly. Generally, it can be seen that the ingroup rejection rate is high fluctuating around an average of 51% while the ingroup confusion rate fluctuates around 3.7%. This compares with the ingroup rejection rate of 19.2% and ingroup confusion rate of 1.0% obtained after training is complete (round 8). The high ingroup error rates for training can be explained as follows. First, the very first training messages for each caller must be either rejected or confused since no model exists for the caller. Of the 160 training messages in rounds 1 to 7, 37 fall in this category. Thus, the combined rejection and confusion rate can never be less than approximately 23%. In the experiment, 34 of these (21.2%) are rejected and 3 (1.9%) are confused. Second, until maturity, the models for some callers are not yet sufficiently representative of the caller and are more susceptible to rejection and confusion. The outgroup accept rate generally rises slowly to 5% with successive rounds as more and more caller models are trained.

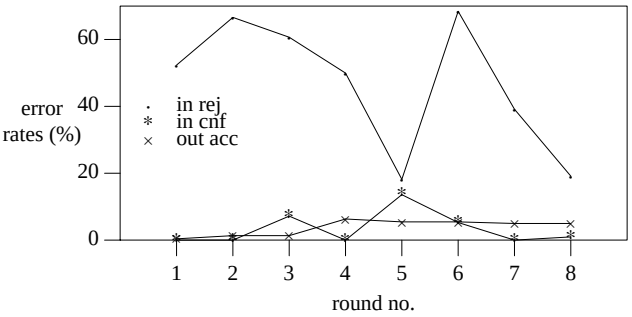

Figure 2. Error rates for 7 successive training rounds (see text); error rates for round 8 are for test messages after training for all callers is complete.

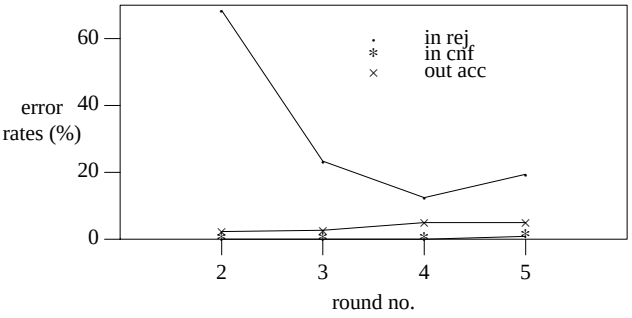

Figure 3. Error rates for 4 successive training rounds (see text); error rates for round 5 are for test messages after training for all callers is complete.

In a contrasting scenario, illustrated in Fig. **??**, all 20 caller models are trained to stage 1 in round 1, to stage 2 in round 2, etc. up to maturity in round 4. In round 1 there are no caller models, so that the ingroup reject rate is 100% and ingroup confusion and outgroup accept rates are 0.0. These results are not shown. In round 2, the stage 2 training messages and the outgroup messages are tested against the stage 1 models. The ingroup reject rate is about 68% while the ingroup confusion rate is 0.0. In successive rounds, the ingroup reject rate drops as the models become more representative while the ingroup confusion rate remains at 0.0. The outgroup accept rates climbs slowly to 5% as the ingroup models become less specific. Actual training scenarios are likely to be some combination of the two described here.

5. DISCUSSION AND CONCLUSION

By definition of the application, voicemail foldering by callers is a challenging task because it is an open-set identification problem. Error rates, as shown in Table 1, must increase as group size increases. Moreover, the application requires that misidentification be kept as low as possible. Misidentification is chiefly attributable to outgroup acceptance since ingroup confusion is generally small, and often insignificant depending on conditions. Maintaining misidentification at a low level means that ingroup rejection can be quite high. Another challenge for the application is the variety of channel and recording conditions that can be expected. Among the features that have been shown to improve performance, particularly with respect to variable conditions, are mel filter bank cepstrum front end, multiple background models, speaker dependent adaptive thresholds, and adapting models using rejected messages.

The application also has some intrinsic advantages compared to, for example, speaker identification used for access control and security applications. The most significant advantage is that it is relatively easy to supervise the training and updating of caller models and to correct identification errors with the cooperation of the subscriber. Thus, rejected messages, messages labelled as "unknown", once the subscriber listens and labels them, serve to create, extend, and update caller models and in the process make the models more representative of the channel and recording conditions associated with each caller. Even misidentified messages can be corrected by the subscriber once they are listened to. When rejected messages are used to update caller models, the best overall performance is obtained, an 11% or 12% ingroup reject rate with outgroup acceptance held at 2.7%.

It should be noted that, at the moment, there are application parameters which could significantly impact performance for which little information is at hand. For example, we have chosen ingroup folder size to range from 1 to 20, but we cannot be sure what would be typical and/or useful. We also do not know what the *a priori* probability of an ingroup message is relative to an outgroup message.

Although not addressed in this paper, other sources of information can be used to provide caller information, such as automatic number identification (ANI) which is widely available in the U.S., and spotting the caller's name and/or number in the message using speech recognition. Combining all these source of caller information should provide a highly robust and useful service for voicemail subscribers.

REFERENCES

[1] 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.

[2] 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.