TRANSCRIPTION AND SUMMARIZATION OF VOICEMAIL SPEECH

Konstantinos Koumpis and Steve Renals

Dept. of Computer Science, University of Sheffield
Regent Court, 211 Portobello St. Sheffield S1 4DP, UK
{*k.koumpis,s.renals*}*@dcs.shef.ac.uk*

ABSTRACT

This paper describes the development of a system to transcribe and summarize voicemail messages. The results of the research presented in this paper are two-fold. First, a hybrid connectionist approach to the Voicemail transcription task shows that competitive performance can be achieved using a context-independent system with fewer parameters than those based on mixtures of Gaussian likelihoods. Second, an effective and robust combination of statistical with prior knowledge sources for term weighting is used to extract information from the decoder's output in order to deliver summaries to the message recipients via a GSM Short Message Service (SMS) gateway.

1. INTRODUCTION

As the emphasis in cellular networks changes from voice-only communication to a rich combination of content based applications and services, speech recognition can provide access to several types of information through a number of portable solutions, including mobile phones and personal digital assistants. This paper deals with the problem of realizing a system that automatically delivers personalized content from voicemail systems to hand held terminals – especially for messages recorded by answering services other than the one provided by the network operator.

Users of existing systems on the receipt of a voicemail notification have to call their answering service and listen to their messages. However, a summary of spoken messages can be proactively[1] delivered as text on a mobile display without the need to call back. The integration of speech recognition and SMS is adequate for capturing and distributing information quickly, no matter the location and without human intervention. Other advantages include uninterrupted information flow in noisy places (crowded streets, train stations, airports) or in so called 'mobile phone free' environments (conferences, meetings), better message management (visual listing and indexing of messages) and lower cost of receiving calls while roaming abroad.

The rest of the paper is organized as follows: in sections 2 and 3 we describe the Voicemail corpus and the experimental setup of the recognizer. The text analysis for the summarization purposes and the evaluation framework are presented in section 4, while the paper is concluded in section 5.

2. VOICEMAIL CORPUS

The system presented herein is trained using the 14.6 hours of speech contained in the Voicemail Corpus Part I (distributed by the LDC) and we refer to this set as VMail15. This corpus was collected from volunteers at various IBM sites in the United States, and comprises 1801 messages in the training set and 42 messages in the development test set. In our implementation the first 1601 messages are used as a training set and the remaining 200 as a validation set.

Voicemail speech is characterized by a variety of speaking rates, accents, tasks and acoustic conditions [6]. Additionally, phenomena such as disfluencies, restarts, repetitions and broken words are common. Another feature of this corpus is that speakers do not receive any direct feedback when they leave messages. This leads to many questions and instructions, which are absent from read or conversational speech. The telephone channel also poses problems of low bandwidth and signal-to-noise ratio as there are no restrictions in location or type of phone used to leave a voicemail message, while some degradation is due to the file compression method used by voicemail systems.

3. EXPERIMENTAL SETUP

The transcription of the 14.6 hours of voicemail data contains approximately 150K words. The basic vocabulary, derived from the one used in the 1998 ABBOT Broadcast News (BN) transcription system, contained an average of 1.3 pronunciations for each of the 65K word vocabulary. The final voicemail vocabulary contained 10K entries. There were about 1K out of vocabulary (OOV) words that were constructed manually following subword pronunciation rules. Several OOV entries were due to transcription errors. The OOV rate of the test set with respect to the final vocabulary was 7.3%.

In all the experiments reported herein we use a hybrid system that combines the temporal modeling capabilities of hidden Markov models (HMM) with the pattern classification capabilities of multilayer perceptrons (MLP). In such a system, a Markov process is used to model the basic temporal nature of speech signal, while the MLP is used as the acoustic model within the HMM framework. The MLP takes acoustic features as an input and estimates *a posteriori* context-independent phone class probabilities [1]. A nine frame window centered on the frame of interest was used as an input to our MLP networks that have a single sigmoidal hidden layer of 2,000 units and an output of 54 phoneme classes[2]. The number of parameters for each layered network was $((9\times\text{input vector length})+54)\times 2000$ weights, plus $(2000+54)$ biases. The

[1] The notion of this service is that the content is delivered to the mobile terminal directly from a third party source without an explicit user request. Short Messages (SMs) within the GSM are transmitted over the mobile phone's air interface using the signaling channels so there is no delay for call setup. SMs are stored by an entity called Short Message Service Centre (SMSC) and sent when the recipient connects to the network. The Wireless Application Protocol (WAP) over SMS supports segmentation and reassembly of SMs allowing the development of more sophisticated services.

[2] In the current implementation no separate phone classes for non-speech phenomena such as 'click' or 'mumble' that are common in voicemail data were created.

System configuration	*WER%*
(1) BN Acoustics, VMail15 bigram	67.0
(2) VMail15 Acoustics, VMail15 bigram	56.1
(3) Combination of (1) and (2), VMail15 bigram	55.2
(4) Embedding training of (3), VMail15 bigram	54.2
(5) VMail15, BN bigram 100:1 weighted mixture	54.4
(6) VMail15, BN bigram 50:1 weighted mixture	53.9
(7) VMail15, BN bigram 20:1 weighted mixture	54.4
(8) VMail15, BN bigram 1:1 weighted mixture	58.0

Table 1: Recognition performance after bootstrapping from the BN acoustics. Impact of combinations of the VMail15 and BN bigrams on recognition accuracy.

input vector length was 13 or 28, depending on the feature extraction technique employed (sections 3.1 and 3.4). We use Viterbi training, with the network parameters estimated using stochastic gradient descent. The outputs were generated by a softmax function computed from the weighted hidden unit outputs. The underlying statistical model was an extremely simple HMM. For each of the 54 phonetic classes, we had an HMM consisting of strictly left-to-right model with multiple states tied to a single distribution; multiple repeated states were used to establish a minimum duration of each phone. The emission probabilities of the HMM were scaled likelihoods estimated by dividing the network outputs by the priors of each class while the transition probabilities were set to 0.5.

3.1. Preprocessing and bootstrapping

The speech signal was segmented into 32 ms frames with a 16 ms frame step. For the feature extraction we created 12th order Perceptual Linear Prediction (PLP) [3] cepstral coefficients plus the log energy (13 elements in total). Feature vectors for a given message were normalized according to the mean and variance of each message in the training data. In order to produce the labels for our initial system we passed Voicemail data through a network trained on 8 KHz bandlimited BN speech with a word error rate (WER)[3] of 36.0% on that task [8]. That system yielded a 67.0% WER for VMail15 as shown in Table 1. We then trained the MLP network with the acoustics of VMail15 and tested it using a bigram language model to get a WER of 56.1%. By combining acoustic probability streams framewise in the log domain the WER came down to 55.2%. The next step was to perform an embedded training of that system achieving a WER of 54.2%.

3.2. Language Model

We also experimented with weighted mixtures of VMail15 language model probabilities with a BN data transcription set containing 1M words. Although we got a slight improvement of 0.3% absolute after combining voicemail and BN data with a ratio of 50:1, we shortly abandoned the use of combined language models, not wishing to increase the vocabulary up to 35K words with a marginally useful technique.

3.3. Multi-words

Since the pronunciation of a word depends on contextual factors such as following and preceding words, word predictability and speaking rate, we also examined ways to add more contextual influence into the pronunciation model. Our initial model consisted of 193 multi-word baseforms which are marked in the transcriptions and were used as single lexical items in both the vocabulary and the *n*-gram language models. We modeled their pronunciations using rules described in [2]. When the baseline vocabulary was augmented with multi-words a small reduction in WER (0.5% absolute) was gained using a bigram model. Since in that set there were included infrequently occurring word-pairs (which would have limited impact on any WER statistics), an additional constraint of word pair frequency was added. After setting a threshold of 15 or more occurrences in order to include words in multi-words list, we came up with 53 multi-words and the respective model produced a WER of 52.4%. However, when the same multi-words were incorporated into the trigram language model this gain vanished as shown in Table 2. Reduced performance can result from crude pronunciation models which seem to assist language models of lower complexity in making better assumptions. Further study is necessary to examine whether the above results are associated with the context-independent nature of our system.

System configuration	*WER%*
(9) 193 multi-words in (4), VMail15 bigram	53.7
(10) 53 multi-words in (4), VMail15 bigram	52.4
(11) PLP alone, VMail15 trigram	51.2
(12) MSG alone, VMail15 trigram	51.8
(13) PLP+MSG, VMail15 trigram, 53 multi-words	48.4
(14) PLP+MSG, VMail15 trigram, no multi-words	46.8

Table 2: Effect of modeling multi-words in performance and comparison of different acoustic features.

3.4. MSG features

We also used Modulation Filtered Spectrogram (MSG) features which provide a robust representation in adverse acoustic conditions. MSG is based on two signal processing strategies modeled after human speech perception [4]. Firstly, changes in the spectral structure of the speech signal (measured with critical-band-like resolution) occurring at rates of 16 Hz or less are emphasised. Secondly, adaptation to slowly-varying components of the speech signal is implemented as a form of automatic gain control.

Although MSG features alone were not as good as PLP features, the WER was significantly reduced by combining these two subsystems as Table 2 depicts. MSG features offered a significant benefit for messages that were degraded in some manner, and were therefore used as the basis of our subsequent work.

4. SUMMARIZATION OF VOICEMAIL MESSAGES

The SMS as a message transmission mechanism introduces an essential limitation on the amount of characters that each message can deliver. As the maximum length of each SM is 140 octets (sufficient coding for 160 7-bit ASCII characters), we need an effective summarizer capable of distilling only the most important information from a message. In theory, we need to generate summaries that use only information that is not in error. Since we can never guarantee that we can perfectly identify the words in decoded audio that are not in error, we cannot depend on sentence-level parsing because even a single incorrect word can completely garble syntactic structure. Instead, we need to choose informa-

[3] The NIST standard scoring package “sclite” was used in all experiments.

tion in another way, one that does not depend on syntactic or semantic analysis. In [10] it has been demonstrated that a combination of confidence measures with simple information retrieval (IR) and information extraction (IE) techniques can be used to accept/reject words and! /o! r! phrases for inclusion in summaries.

In our work the task of message summarization may be cast as a combination of message statistics and prior knowledge, where the decision to accept or to reject words or phrases in a summary is based upon speech recognition confidence measures, term collection frequency and named entity (NE) lists.

Each message is represented as a vector of weighted terms. The idea behind term weighting is selectivity: what makes a term an appropriate one is whether it can express correct information content from the original message.The computation of the weights reflects empirical observations of the data. The algorithm depicted in Figure 1 removes the words with the lowest score given the available message length which is determined by the level of compression. A description of the three main weight factors along with a lossless summarization technique follows. The impact of each weight factor can be empirically optimized, but as a general rule we consider confidence measures and NE to be highly important.

Confidence measures quantify how well a model matches some spoken utterance, where the values must be comparable across utterances. Hybrid connectionist systems are well suited to producing computationally efficient acoustic confidence measures [11]. A discriminating confidence measure may be obtained using a duration normalized sum of log posterior probability estimates. For a phone q_k which a hypothesized start time n_s end time n_e given some acoustic data x^n the confidence measure is:

$$CMW_{npost}(q_k) = \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} \log\left(p(q_k|x^n)\right) \qquad (1)$$

These confidence measures are computed directly by CHRONOS decoder [7] where the language model is used to constrain the search for the optimal state sequence but is not used in the computation of the confidence estimates.

Collection Frequency is inspired from IR and is based on the fact that terms which occur only in a few messages are often more likely to be relevant to the topic of that message than ones that occur in many. For a term t_i the collection frequency is defined as:

$$CFW_{t_i} = log \frac{N}{n_{t_i}} \qquad (2)$$

where N is the number of messages in the training data and n_{t_i} is the number of messages that term t_i occurs in. The CFW_{t_i} weights are then normalized to the number of messages.

Named entity lists were employed in order to prioritize words that may be classified as proper names, or as certain other classes such as organization names, dates, times and monetary expressions. This is less straightforward than identifying NE in written text, since speech recognition output is missing features that may be exploited by "hard-wired" grammar rules or by attachment to vocabulary items, such as punctuation, capitalization and numeric characters. Our NE lists constitute of 3.4K entries, 2.8K of which derived from the BN corpus [9] whereas the remaining were classified manually from the VMail15 transcriptions. This allowed us to retain in the summaries certain types of the terms containing important information, i.e. series of digits comprising telephone numbers or proper names.

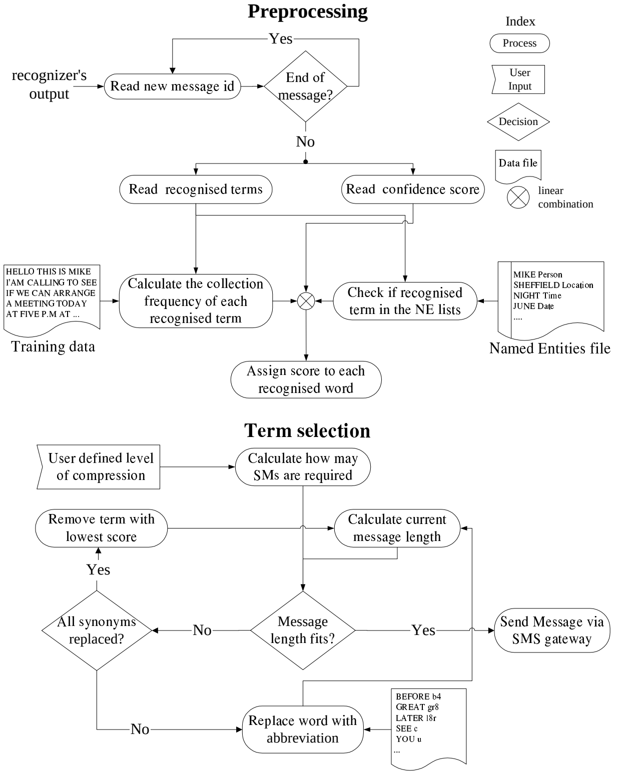

Figure 1: A schematic outline of the mechanism that summarizes the decoder's output using a combination of statistics and prior knowledge.

Abbreviations and digits is another way to reduce the length of the message without losing information and is based on the fact that users of SMS are familiar with text messaging abbreviations. As an example the phrase "SPEAK TO YOU LATER" can be replaced with "spk 2 u l8r". In our system we have been using a set of approximately 40 such abbreviations that offer a reduction of more than 10% in the message length. We also replaced words describing numbers with the respective digits, i.e. the word "THREE" was replaced by "3" in the summaries offering a further reduction in length and increasing message readability.

4.1. Evaluation of summaries

Summaries are inherently difficult to evaluate because the quality of a summary depends both on the use for which it is intended and on a number of other human factors, such as how readable an individual finds a summary or what information an individual thinks should be included in a summary. In order to evaluate how the summaries reflect the information content of the original messages of the Voicemail task, two complementary methods were employed. First, a web-based survey completed by volunteers using a five-point scale (1=Poor, 5=Excellent) and second, the recently proposed Slot Error Rate (SER) [5] was calculated.

Figure 2 shows the average score of summaries of the 10 longest messages of the test set for two different levels of compression[4] as judged by 14 subjects after comparing with the original transcriptions. Some low scores could be explained by the fact that 5 of the subjects were not familiar with SMS abbreviations. Although further validation with larger test set is necessary, these initial message summaries were considered usable, as the results correspond to 3.14 and 2.56 scores for a compression level of 40% and 20% respectively.

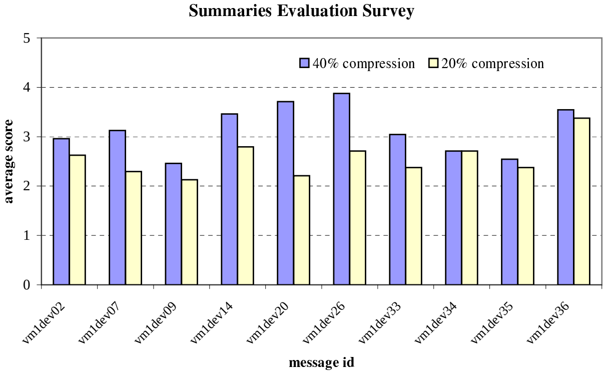

Figure 2: Survey results using a five-point scale on how well each summary retains the information contained in the original transcriptions of the 10 longest messages of the test set.

As a statistical measure of evaluation, SER was selected as more applicable for this task due to the relatively short length of voicemail messages. This is analogous to the WER and does not have the drawbacks of F-measure, which is computed by the uniformly weighted harmonic mean of precision and recall [5]. SER is equal to the sum of the three types of errors – substitutions S, deletions D and Insertions I – divided by the total number of slots in the reference:

$$SER_{msg} = \frac{S+D+I}{C+S+D} \tag{3}$$

We define a slot to be any term or group of terms containing key and essential information for the message recipient. The baseline SER for the 42 messages test set was 40.3%. For a 40% compression level the SER of the summaries was 51.7%, while for a 20% compression (corresponding to one SM per voicemail message) the SER raised to only 55.9%. This indicates that the SER is mainly due to the transcription errors and the 7.3% OOV rate, rather than the summarization approach. From these initial experiments is shown that the statistical model in combination with prior knowledge sources is an effective and robust approach to message summarization.

5. CONCLUSION

We have described a system that transcribes and summarizes voicemail messages contained in the LDC Voicemail corpus. The initial results on the transcription task demonstrate that competitive performance can be achieved using a fraction of parameters than those required by systems based on mixtures of Gaussian likelihoods. Although BN data did not offer significant gains into the language models, we benefited from bandlimited acoustic data and the NE lists derived from this task. Finally, the summarization mechanism that we have presented has a simple and explicit link to the models, allowing extraction of subtle information regardless of the number of transcription errors and the relatively high OOV rate.

6. ACKNOWLEDGEMENTS

We gratefully acknowledge Dan Ellis for providing the bandlimited BN trained MLP and his assistance on the front-end setup, Yoshi Gotoh for his help with the language modeling and Mark Stevenson for providing the NE lists from the BN corpus. Additionally, we thank Dave Abberley, Rob Gaizauskas, Tony Robinson and Gethin Williams for valuable discussions during this work. Thanks also to the volunteers who participated in the summaries evaluation survey.

7. REFERENCES

1. H. Bourlard and N. Morgan, *Connectionist speech recognition: A hybrid approach*, Kluwer Academic Publishers, 1994.
2. M. Finke and A. Waibel "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," *Proc. of EuroSpeech*, vol. 5, pp. 2379-2382, Rhodos, Greece, 1997.
3. H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *Journal of the Acoustical Society of America*, vol. 87, pp. 1738-1752, 1990.
4. B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram", *Speech Communication*, vol. 25, pp. 117-132, 1998.
5. J. Makhoul, F. Kubala, R. Schwartz and R. Weischedel, "Performance measures for information extraction", *Proc. of the DARPA Broadcast News Workshop*, pp. 249-252, Virginia, USA, 1999.
6. M. Padmanabhan, E. Eide, G. Ramabhardan, G. Ramaswany and L. Bahl, "Speech recognition performance on a voicemail transcription task", *Proc. of the ICASSP*, vol. 2, pp. 913-916, Seattle, USA, 1998.
7. T. Robinson, J. Christie and G. Cook, "Time-first search for speech recognition", *Submitted to Speech Communication*, 2000.
8. T. Robinson, G. Cook, D. Ellis, E. Fosler-Lussier, S. Renals and G. Williams, "Connectionist speech recognition of Broadcast News", *Submitted to Speech Communication*, 2000.
9. M. Stevenson and R. Gaizauskas, "Using corpus-derived named lists for named entity recognition", *Proc. of Applied NLP and the N. American Chapter of the ACL*, pp. 290-295, Seattle, USA, 2000.
10. R. Valenza, T. Robinson, M. Hickey, and R. Tucker, "Summarization of spoken audio through information extraction", *Proc. of the ESCA Workshop on Accessing Information in Spoken Audio*, pp. 111-116, Cambridge, UK, 1999.
11. G. Williams and S. Renals, "Confidence measures from local posterior probability estimates", *Computer Speech and Language*, vol. 13, pp. 395-411, 1999.

[4] The ratio of summary length to source length in characters including spaces between words.

[5] Precision deals with S and I errors and recall with S and D errors.