International Conference on Spoken Language Processing, pp. 2011-2014, ISCA, Pittsburgh, PA, USA, 2006

Call Analysis with Classification Using Speech and Non-Speech Features

Yun-Cheng Ju, Ye-Yi Wang, and Alex Acero

Microsoft Corporation
One Microsoft Way, Redmond, WA 98052, USA
{yuncj, yeyiwang, alexac}@microsoft.com

Abstract

This paper reports our recent development of a highly reliable call analysis technique that makes novel use of automatic speech recognition (ASR), speech utterance classification and non-speech features. The main ideas include the use the N-Gram filler model to improve the ASR accuracy on important words in a message, and the integration of recognized utterance with non-speech features such as utterance length, and the use of utterance classification technique to interpret the message and extract additional information. Experimental evaluation shows that the use of the utterance length, recognized text, and the classifier's confidence measure reduces the classification error rate to 2.5% of the test sets.
Index Terms: Call Analysis, Answering Machine Detection, ASR, N-Gram Filler, Utterance Classification

1. Introduction

Various outbound telephony applications require *call analysis* for identifying the presence of a live person at the other end. A typical scenario is the outbound call management (OCM) system in an outbound call center (a typical telemarketing center) where agents can increase their efficiency by only attending the calls which are identified as answered by live persons. Another emerging use of the OCM systems is for outbound telephony notification applications where, for example, an airline company can notify their customers for flight delay or changes. The ability of robust call analysis has been one of the few top requests from the Microsoft Speech Server customers.

Prior to the work presented in this paper, most work in this area have been focusing on two components - a **tone detector** for busy, non-existing numbers, Modem, or Fax and a **voice discriminator** for human voices. Tone detection [1] can be done accurately for all three types of tones (Special Information Tones (SIT), Cadence Tones, and Modem/Fax Tones). However, voice discriminator is more difficult and a typical "*fast voice*" approach [2] is to use the zero-crossing and energy-peak based speech detector algorithm with the silence and energy timing information presented in the signal. It has been observed that people usually answer the phone by saying a short "Hello" (or an equivalent greeting in other languages) and wait for the other party to respond so there is an energy burst followed by a longer period of silence. On the other hand, the recorded voice (answering machine) tends to be speech signals with normal sequence of energy and silence periods.

With the advent of new technology, many of the traditional audible tones have been augmented by synthesized and recorded human speech. For example, the "non-existing phone numbers tone" has largely been replaced by recorded speech. This trend has made the traditional tone detectors increasingly less effective. In addition, phone screening technology, fueled by the increasing number of unsolicited telemarketers, also poses new challenges for the voice discriminators to let legitimate automated calls (e.g., the notification service) to complete.

Recently, a few ASR-based approaches have been proposed to improve the voice discrimination part of the problem [4]. However, high word error rates have limited its success.

In this paper, we focus our efforts on **voice discriminator** and present an improved technique over that described in [2] to overcome the hurdles in using ASR. We use the N-Gram based filler model [3] from limited training sentences to dramatically improve word accuracy for those *information bearing* words in users' utterances. We use speech utterance classification to overcome fragility of the keyword spotting approach by analyzing the whole utterances and perform concept recognition based on automatically learned features from training sentences. In addition, we also use non-speech features, mostly the utterance length, to form a prior to further improve the classification accuracy and to provide a safe fall back for failed recognitions. Finally, our speech utterance classifier's capability of extracting critical information (e.g., "press 2" in "… if you are not a telemarketer, **press 2** to complete the call") enables legitimate applications to pass the phone screening or other automated systems.

The rest of the paper is organized as follows. Section 2 briefly reviews the N-Gram based filler model technique and describes the recognition language model. Section 3 summarizes the utterance verification technique and our classification model. Section 4 describes the integration of the classification with non-speech features and evaluates the proposed methods. Section 5 concludes the paper.

2. N-Gram Based Filler Model

The integrated technique is based on our earlier work on N-Gram based filler model [1], which allows us to reduce both the recognition error rate and the training corpus size. In this section, we briefly review our model and the related *interpolated* approach.

2.1. Interpolated N-Gram Model

It is well known that statistical language model (SLM) increases robustness of speech recognition. It is also well known that an SLM can be implemented as a Context Free Grammar (CFG) [5].

There are several approaches in using SLM for ASR. A **general-domain** N-Gram (typically used for desktop dictation) can be used without development effort, but its performance is mediocre. A **domain-specific** LM (completely trained from in-domain training data) works best, but it sometimes may not be practical due to the requirement of a large amount of domain-specific training data. A **domain-adapted** LM, on the one hand, works better than the general-domain LM and requires fewer training sentences. However, the size of an adapted LM, which is no smaller than the general-domain N-Gram, is prohibitive for dialog turn specific language modeling on a speech server.

We proposed an **interpolated approach** where a domain-specific core LM is trained completely from a limited set of training sentences. At the run time, a general domain garbage model **N-Gram Filler** (typically a trimmed down version of the dictation grammar) is attached through the unigram back-off state for improved robustness. The interpolation is achieved by discounting some unigram counts and reserving them for unseen garbage words. The size of the resulting LM is small (typically a few Kilo bytes) due to the small size of the training sentences required. The N-Gram Filler is relatively large (8 MB in our system) but poses no harm to the speech server because it is shared among all dialog turns and different applications. Figure 1 depicts the binary CFG trained from a single training sentence "press <digit> to complete …". In addition to the three special states (<S>, </S>, and unigram backoff state), states with one word (e.g. "press") are bigram states and states with two words (e.g. "press <digit>") are trigram states.

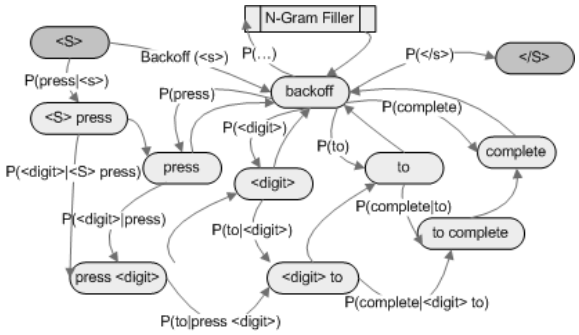

Figure 1: Interpolated N-gram Model. (Transitions w/o labels are back-off transitions)

2.2. Accuracy Improvement for Information-Bearing Words

For the concept recognition application described above and other speech utterance classification applications, improvement of word accuracy may be too costly and unnecessary. The preliminary experiments showed that our interpolated N-gram model can effectively recognize *important* (i.e., information bearing) words correctly without acquiring massive training sentences. We believe the improvement in word accuracy for information bearing words is the key to lower the development cost of speech applications.

For example, the LM depicted in Figure 1 recognizes the utterance "In order to complete the call, please press 2 now" as "feel the **to complete** a call Steve **press 2** know". Even though the word accuracy rate is unimpressively at 50%, all of the information bearing words, as occurred in the training sentence, were correctly recognized.

3. Speech Utterance Classification

As perfect speech recognition of either live person or answering machine is unrealistic, using speech utterance classification significantly improves the robustness over techniques based on keyword spotting approaches [6,7]. In this section, we first briefly summarize the use of maximum entropy (Maxent) classifier [8] and then describe the heuristic approach we used to integrate some non-speech features.

3.1. Maximum Entropy Classifier

Given an acoustic signal A, the task of speech utterance classification is to find the destination class $\hat{C}$ from a fixed set $\mathscr{C}$ that maximizes the conditional probability $\mathrm{P}(C \mid A)$:

$$\begin{aligned}\hat{C} &= \arg\max_{C\in\mathscr{C}} \mathrm{P}(C \mid A) \\ &= \arg\max_{C\in\mathscr{C}} \sum_{W} \mathrm{P}(C \mid W, A)\mathrm{P}(W \mid A) \\ &\approx \arg\max_{C\in\mathscr{C}} \sum_{W} \mathrm{P}(C \mid W)\mathrm{P}(W \mid A) \\ &\approx \arg\max_{C\in\mathscr{C}} \mathrm{P}\left(C \mid \arg\max_{W} \mathrm{P}(W \mid A)\right)\end{aligned} \tag{1}$$

Here W represents a possible word sequence that is uttered in A. We made a practical Viterbi approximation by adopting a two-pass approach, in which an ASR engine obtains the best hypothesis of W from A in the first pass; and a classifier takes W as input and identifies its destination class. The text classifier models the conditional distribution $\mathrm{P}(C \mid W)$ and picks the destination according to Eq. (1). The Mexent classifier models this distribution as

$$\mathrm{P}(C \mid W) = \frac{1}{Z_\lambda(W)} \exp\left(\sum_{f_i\in\mathscr{F}} \lambda_i f_i(W, C)\right) \tag{2}$$

Where $\mathscr{F}$ is a set that contains important features (functions of a destination class and an observation) for classification, and $Z_\lambda(W) = \sum_C \exp\left(\sum_{f_i\in\mathscr{F}} \lambda_i f_i(W, C)\right)$ normalizes the distribution. We include binary unigram/bigram features in $\mathscr{F}$:

$$f_{u,c}(W, C) = \begin{cases} 1 & \text{when } u \in W \wedge C = c \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where u is a word or a word bigram, c is a destination class.

In Eq. (2), λ_i's are the weights for feature f_i, which are optimized from training data.

Finally, we define the confidence measure as the posterior ratio:

$$Conf(C \mid W) = \frac{P(C \mid W)}{P(C' \mid W)} \tag{4}$$

where C' is the class that has the highest posterior probability among those different from C.

3.2. Incorporating Non-Speech Features

There are many useful non-speech features we can obtain from ASR (e.g., the length of utterance and initial silence). In fact they had long been used for this application. It is important to find an efficient and effective way to incorporate those features for better performance. In addition, these features also provide practical backup strategies when the ASR fails occasionally. One such a feature is utterance length. We have observed that the responses of live persons are mostly shorter than the answering machine messages.

One possible way of incorporating those features is to treat them as features in a different dimension and to form joint features. However, this approach requires more training data.

We took a simpler, heuristic two-pass approach. When the response is short, we assume it's from live person unless there is strong evidence otherwise from the word level classifier and vise versa for the long responses. The algorithm is illustrated in Figure 2. The "Very Short", "Very Long", and "Relatively Short" are set to be 1, 4, and 2.5 seconds observed from the training set.

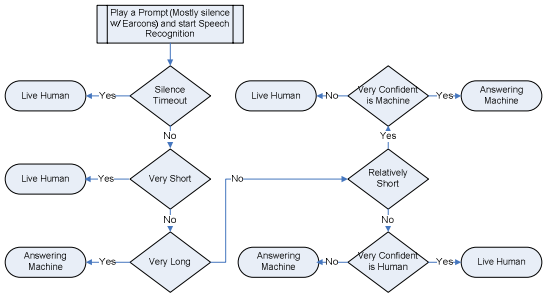

Figure 2: Call Classifier using both utterance length and recognized text

As will be shown in the results section, this simple heuristic significantly improved the performance by recovering the classification errors from those poorly recognized sentences.

4. Experiments & Results

We built a prototype system on a pre-release version of Microsoft Speech Server 2007 to initiate outbound calls. The training samples consists of 1331 connected calls to residential numbers across the United States. (Table 1).

Live Person (508)	Hello (420)
	Other Greetings (88)
Recorded Message (823)	Answering Machine (809)
	Failed / Invalid Numbers (7)
	No Solicitation Screening (7)

Table 1. Breakdown of the training set in each category

We have two test sets. Test set #1 (Res) is from the 2629 successfully reached residential numbers different from those in the training set, Test set #2 (Mix) is from the connected calls to 2135 mixed residential and business numbers. The second test set contains more varieties and is more challenging. There was no overlapping among the three data sets.

The evaluation is conducted for two tasks. The first task is to distinguish a live person from the recorded message. The other task is to subdivide the recorded message into one of the three categories (Answering Machine, Failed Calls, and No Solicitation Screening) and to extract the action necessary to complete the call in the case of phone screening.

From the training samples, we created a training set of 90 sentences/phrases in a total of 1061 words (mostly for recorded messages) for both the recognition grammar and call classifier training.

4.1. Detection of Live Person

In the first experiment, we investigated the effect of using both speech and non-speech features for the detection of live person vs. recorded message as described in section 3.2. Table 2 shows the break down statistics of the test set according to the utterance duration (D) or the text classifier's output (T).

Truth	Classification	Res.	Mix.	Both
Live Person	D: Very Short	955	713	1668
	D: Short	76	298	374
	D: Long*	6	35	41
	D: Very Long*	1	3	4
	T: No Recognition	334	335	669
	T: Live	603	494	1097
	T: Recorded*	101	220	321
	Total	1038	1049	2087
Recorded Message	D: Very Short*	2	0	2
	D: Short*	113	62	175
	D: Long	209	109	318
	D: Very Long	1267	915	2182
	T: No Recognition*	44	27	71
	T: Live*	32	23	55
	T: Recorded	1515	1036	2551
	Total	1591	1086	2677

Table 2. Break down classification results of the test sets. The rows marked with (*) represent misclassifications if only the durations or text classification is used alone. Our proposed method improves the cases in the shaded area.

It is observed that using the recognized utterance alone for the call classification yields the worst result. The overall error rate is around 9.4% if we treated the cases of no recognition as live persons ((321 + 71 + 55) / (2087 + 2677)). It is even less accurate than only using the utterance length where the error rate is 4.7% ((41 + 4 + 2 + 175) / (2087 + 2677)). More training samples might be needed to reduce the recognition errors to reduce the classification error rate.

In our proposed approach, we use the utterance length as the first cut and then use the call classification technique for the cases in the middle range (shaded areas in Table 2).

Table 3 shows our approach is very effective for short utterances. Only the shaded areas are errors. We believe that is because the short messages are more precise (e.g., "Please leave your message") and are also easier to get recognized and classified correctly.

The error rates are summarized in Table 4. We found that we have reached some small business receptionists and their reserved phone numbers in the second (Mixed) test set which there is a slight mismatch to our training set. However, the

combined overall error rate (2.5%) is very appropriate for most telephony applications.

Truth	Classifier's outcome	Res.	Mix.	Both
Short Live	Very Confident is Machine	5	41	46
	Otherwise	71	257	328
Long Live	Very Confident is Human	1	9	10
	Otherwise	5	26	31
Short Recorded	Very Confident is Machine	92	52	144
	Otherwise	21	10	31
Long Recorded	Very Confident is Human	3	4	7
	Otherwise	206	105	311

Table 3. Classification results from our proposed approach. The shaded rows represent errors.

	Res.	**Mix.**	**Both**
Utterance Classification Only	6.7%	12.6%	9.4%
Duration Only	4.6%	4.7%	4.7%
Proposed Approach	1.4%	3.9%	2.5%

Table 4. Overall classification error rates.

4.2. Sub-Classification of Recorded Messages

One advantage of our ASR/speech classification approach is the capability of understanding the message to keep up with new telephony/speech technology, like the use of recorded speech in the telephony infrastructure and in the phone screening/auto-attendant systems.

In this experiment, we demonstrate the performance of the proposed approach in classifying the recorded messages into sub-classes and extracting the actions needed to pass the "No Solicitation" screening systems. We built a 3-way utterance classifier with the same training sentences described earlier. Table 5 shows the classification results on the recorded messages in the test sets.

	Res.			Mix.			Both		
	M	F	N	M	F	N	M	F	N
Machine (M)	1564	4		1017	3		2581	7	
Failed (F)	1	11		11	51		12	62	
No Solici. (N)	2		9			4	2		13

Table 5. Confusion matrix for classifying the recorded message into three sub-categories. Numbers in the shaded blocks are errors.

6 out of the 11 misclassified "Failed Call" messages in the second (Mixed) test sets were not generated by the carrier and the key phrases like "unassigned number", "call again later for further information", "non-working", "temporarily out of service" were not covered in our limited training sentences. The rest of errors were caused by misrecognition. We believe more training data can further improve our performance. In addition, the action needed to pass the No Solicitation screening system for the 13 cases were all correctly extracted.

5. Summary and Conclusions

Use of ASR to improve call analysis has started to attract more attention in keeping up with the trend of replacing audible tones with recorded messages and the popularity of the phone screening/auto-attendant systems. Yet most practices have not overcome the obstacle of high recognition word error rates. We have extended our earlier work in the N-Gram filler model and used the interpolated SLM to reduce the required amount of the training sentences. This achieves significantly higher accuracy for the *information-bearing* words observed in the training set and contributed to high accuracy in call analysis.

In addition to speech content features, non-speech features are also integrated in the classification framework. Strong classification results are obtained on both call analysis and the new recorded message sub-classification task.

Our integrated approach follows common sense and the facts we observed from the sample calls. Non-speech features, specifically the utterance length, are used in the first stage of classification. Very short utterances tend to come from live persons and very long utterances come from machines. We only perform utterance classifications for the relative short and long utterances with different priors. Only highly confident short answering machine messages are considered as from answering machines and vise versa for highly confident long live person utterances. In this way, a significant number of the recognition and classification errors are avoided.

We have also shown that the use of the N-Gram based SLM interpolation approach worked well in the task of sub-classification of the recorded messages, a typical speech utterance classification + slot filling application. We believe our work opens a new opportunity for two automated telephony systems to interact.

We are currently evaluating the performance of the proposed approach in a bilingual/multi-lingual area without the use of a bi-lingual speech recognizer. Preliminary experiments have shown promising results for call analysis.

6. Acknowledgements

We thank Dong Yu, Li Deng in our group for useful discussions and David Ollason, Craig Fisher in the Microsoft Speech Server group for suggesting the research topic.

7. References

[1] L. J. Roybal, "Signal-Recognition Arrangement Using Cadence Tables" United States Patent, 5,719,932

[2] S. C. Leech, "Fast Voice Energy Detection Algorithm for the Merlin II Auto-Dialer Call Classifier Release 2 (ADCC-M R2)," MT XG1 W60000, 1990.

[3] Dong Yu, Yun Cheng Ju, Ye-Yi Wang, and Alex Acero, "N_Gram Based Filler Model for Robust Grammar Authoring," Proc. ICASSP, 2006.

[4] Sharmistha Sarkar Das, Norman Chan, Danny Wages, and John H. L. Hansen. "Application of Automatic Speech Recognition in Call Classification," Proc. ICASSP, 2002.

[5] G. Riccardi, R. Pieraccini, and E. Bocchieri, "Stochastic Automata for Language Modeling," Computer Speech and Language, Vol. 10, pp. 265-293, 1996.

[6] H.-K. J. Kuo et al., "Discriminative Training for Call Classification and Routing", Proc. ICSLP, 2002

[7] B. Carpenter and J. Chu-Carroll, "Natural Language Call Routing: A Robust, Self-organizing Approach," Proc. ICSLP 1998.

[8] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing," Computational Linguistics, Vol 22, 1996.