Probabilistic Trainable Segmenter for Call Center Audio Using Multiple Features

Nina Zinovieva, Xiaodan Zhuang, Pat Peterson, Joe Alwan, Rohit Prasad

Speech, Language and Multimedia Business Unit
Raytheon BBN Technologies, Cambridge, MA

{nzinovie, xzhuang, ppeterson, jalwan, rprasad}@bbn.com

Abstract
An important component of customer call experience analysis is to distinguish different segments of a call including interactive voice response (IVR), waiting in queue, and interaction with an agent. Because segment information from telephone switches is not always available, or may be difficult to obtain, we sought a method that could perform such segmentation solely from the recorded audio. In this paper, we present a probabilistic framework for segmenting call center audio into IVR, Queue, and Agent using a suite of rich features based on both speech and non-speech content. We study different statistical classifiers such as Maximum Entropy (MaxEnt) and Conditional Random Field (CRF). We present experimental results on real-world call center data and demonstrate that the probabilistic approach achieves superior segmentation performance, and outperforms a rule-based approach, while significantly reducing the time needed to deploy the segmenter for a new call center.

Index Terms: customer call center analysis, segmentation, acoustic events, conditional random field

1. Introduction
Accurate segmentation of audio is a key challenge for automatic speech recognition, spoken language understanding, and other generic audio processing tasks [1][2][3]. A unique segmentation application is in customer call center audio analysis, where the switching information between interaction with agents (Agent), interactive voice response (IVR), and waiting in queues (Queue) provides valuable information regarding the efficiency and effectiveness of the call center. For example, the amount of time spent waiting in the queue to talk to an agent is correlated with user satisfaction. We refer to the task of segmenting call center audio into Agent, IVR, and Queue as high-level segmentation. While such information, in principle, can be obtained from the switch, in practice it is difficult given the logistic constraints such as outsourcing arrangements pervasive in the call center business and privacy consideration.

Existing work in similar high-level segmentation tasks focus on the spoken content of both the customer and the support agent recovered by automatic speech recognition. For instance, Riccardi et. al. [4] performed task-specific ASR decoding and used the automatically identified salient phrases to classify an utterance into call types for routing to an agent pool. Similarly, Support Vector Machines (SVMs) have been used to classify each utterance into agent or customer, and additional call-section types, such as resolution, question, closing, etc. using off-the-shelf ASR system output [5]. Also, content analysis using ASR output has been studied to identify important segments for both the agent and the administrators [6].

While audio recording from both the customer and the call center sides have been successfully utilized in the above work [4][5][6] for call analysis, no literature that we are aware of looks into the feasibility of customer call content analysis using only the call center side recording. Segment information is more salient on the call center side and processing only that half of the call leads to significant cost savings.

This paper focuses on segmenting call-center side only single-channel audio recording into the three non-overlapping categories of Agent, Queue, and IVR, i.e. no customer audio is used in determining the segmentation. The Agent and IVR segments refer to periods when the customer is having a conversation with the agent or the automated system, including times when the agent or the system is listening to the customer. The Queue segments are usually pre-recorded advertisement, music or silence that fills the gap when the customer is put on hold or being transferred to another agent.

We introduce a probabilistic, trainable system based on a comprehensive set of features from speech activity detection, non-speech events analysis, and speech recognition. From a modeling perspective, inspired by successes of discriminative methods [7][8], we explore Maximum Entropy (MaxEnt) and Conditional Random Field (CRF), trained on a small number of manually labeled calls. We demonstrate the efficacy of the statistical system on real-world customer call center data. The proposed system outperforms our baseline, a rule-based expert system, which was previously tuned to the best of our ability for a production application.
2. Overview of the Probabilistic Trainable Segmenter

Figure 1 describes the probabilistic segmenter for call center audio. First, a template-based prompt detector operates on the whole audio identifying portions of the audio that match a set of pre-defined audio prompts. In parallel, we use a speech/non-speech detector to hypothesize non-speech events and speech utterances [12]. Non-speech events include silence/pauses, music, mouth/breath noises, and laughter.

From the speech segments, we detect coarse speech categories corresponding to the target segments types: Agent, IVR, and Queue. In addition, a large vocabulary automatic speech recognition system produces the transcripts for each speech segment. The rich suite of acoustic- and lexical-based features is used by discriminative classifiers to generate the final segmentation hypothesis for the call.

For training purposes, the entire segmenter requires annotation of the target segment types for a small set of calls.

3. Features for Call Segmentation

We first describe the features used for call segmentation. All features are derived solely from the call center side recording, and no customer-side recording is used in this work.

3.1 Acoustic features

3.1.1. Call Center Prompts

Using a correlation-based template matching method, we detect audio menus or messages known to be associated with the queue and the IVR segments in the call center. These detections assign either a prompt-Queue or a prompt-IVR label to each corresponding speech utterance or non-speech segment, referred to as the acoustic prompt features.

3.1.2. Non-speech Events

As illustrated in Figure 1, the speech/non-speech detector generates speech and non-speech events. To do so, we use off-the-shelf HMM-based generalized phoneme class and non-speech event models [12] trained on the Switchboard telephony speech corpora [9]. The Viterbi algorithm, followed by Bayesian Information Criterion (BIC) and bottom-up clustering, is used to segment audio into speech and various non-speech events. The output of the detector for non-speech events is converted into a feature vector of non-speech features.

3.1.3. Coarse Speech Events

We train a set of Gaussian mixture models (GMMs) using the hypothesized speech utterances obtained in Sec 3.1.2, based on the high-level segment annotations of Agent (male or female), IVR, and Queue (speech or non-speech). These GMMs are trained on Mel-frequency Cepstral Coefficients (MFCCs), and hypothesize the above coarse speech events for each speech utterance. We also include the additional feature of speech utterance duration.

3.2 Lexical features

We produce automated transcriptions using a large-vocabulary speech recognizer trained primarily on the Fisher corpus [10]. Speech content on the call center side, when available, is very informative for the problem of high-level segmentation.

We leverage such information by training category-specific n-gram language models (LM). In addition, we select the initial agent utterances (usually related to agent greetings) and the final agent utterances to train two additional language models. We estimate the optimal interpolation weights across the above five language models leading to minimum perplexity on the ASR output for each utterance. This results in five scalar LM-based features for each utterance. Compared with the widely used n-gram features, this approach is more robust when the in-domain data is limited, which is often the case when deploying a call center audio analysis system.
4. Models for call segmentation

We formulate the problem of high-level call segmentation as follows,

\[ Y = \text{argmax } P(Y|X), \]

where \( Y \) is the high-level segmentation of the call, and \( X \) is the collection of features characterizing the call.

Since the various features we consider are associated with speech and non-speech segments in the call, we expand Eq. 1 as,

\[ [y_1 \ldots y_T] = \text{argmax } P(y_1 \ldots y_T|x_1 \ldots x_T), \]

where \( x_i \) and \( y_i \) refer to the features and the high-level segmentation label associated with the \( i^{th} \) segment, \( i \in \{1, \ldots, T\} \). Any adjacent segments with the same hypothesized label are merged.

Recently, MaxEnt and CRF classifiers have been successfully used in segmentation, boundary detection and various tagging tasks [7][8]. In particular, they have been shown to generally outperform traditional hidden Markov models (HMMs) for sentence segmentation [8].

In this work, we investigate the use of these two discriminative approaches for the high-level segmentation task.

First, we tackle the problem as a series of segment classification problems, \( y_i = \text{argmax } P(y_i|x_i), i \in \{1, \ldots, T\} \), based on the \( J \) different features. Each segment classification is solved using the MaxEnt model:

\[ P(y_i|x_i) = \frac{1}{Z} e^{\sum_{j=1}^{J} \beta_j f_j(x_i,y_i)}. \]

The MaxEnt model provides an effective way to encode different features, but does not take into consideration the correlation between different segments. As a result, it is prone to over-segmenting the call.

In contrast, the linear chain CRF jointly models the labels from all segments of a call, therefore avoiding the suboptimal, temporally local decisions made by the MaxEnt model. The CRF can be posed as a sequential extension of the MaxEnt formulation as in Equation 4, where \( Z \) is a normalizer.

\[ P(y_1 \ldots y_T|x_1 \ldots x_T) = \frac{1}{Z} e^{\sum_{j=1}^{J} \sum_{i=1}^{T} \beta_j f_j(x_{i-1},y_{i-1},x_i,y_i)}. \]

5. Experiments

5.1. Experimental Setup

5.1.1. Dataset

The dataset includes call-center-side single-channel recordings of English customer-support telephone calls, where customers interact with IVR (interactive voice response) automated systems or with human agents, and may also spend time listening to music or recorded messages while “in queue” waiting for an agent. The challenge is to determine at any point, using only the audio evidence on the call center side, whether the caller is interacting with an IVR, waiting in queue, or talking with an agent.

We annotated 299 calls by dividing them into segments labeled according to which phase of a call they represented (IVR, Queue, or Agent) and two broad acoustic categories (male or female agent, speech or non-speech queue audio). 199 of these calls are used in training, and the rest of the calls are split into development and test sets, each consisting of 50 calls.

5.1.2 Different feature sets and models

We evaluated the performance of the MaxEnt and the CRF models on the test set. All model parameters were optimized on the development set. We designed feature sets to understand the relative effectiveness of the different non-speech events, coarse speech events, prompt features, and lexical features.

For comparison, we also evaluated the same task using an existing rule-based expert system using the same acoustic features in Sec. 3.1 and a set of manually-crafted rules for combining the utterance-level features into high-level segment labels. The rules are specified in terms of patterns reductions that are applied recursively to sequences of up to 3 adjacent segments. The most common reductions are to:

1. Re-label speech-based segments as IVR, Queue, or Agent as evidence becomes clear,
2. Combine noise or silence into adjacent speech-based segments
3. Merge similar (as defined by an expert) adjacent segments.

The above rule-based system requires substantial effort from a domain expert for each call center deployment.

5.1.3. Metric

For scoring, we used the standard acoustic event detection metric by NIST, i.e., acoustic event detection accuracy (AED-ACC) [11], which is the harmonic mean of the precision and the recall for all the three high-level segment labels. The center of the hypothesized segment should fall into the span of the concerned reference segment, or vice versa, and both labels need to match in order to be considered correct.

5.2. Experiment Results

Figure 2 illustrates a snippet of a call (only the customer call center side), the ground-truth segmentation labels, and features extracted by different system components and the system outputs based on either MaxEnt or CRF.
In Table 1, we first present as baselines the performance of the hand-crafted rule-based expert system as well as a null system, which simply hypothesizes the whole call as a single most frequent segment: “agent”.

Tables 2 shows the performances using the MaxEnt and the CRF models respectively, with various feature sets consisting of (N)on-speech events, (P)rompts, (C)oarse speech events, Lexical (L) features, and speech (D)uration.

We can see that CRF outperforms MaxEnt for all feature combinations. On further examination, we observe that MaxEnt typically produces more segments and excessive transition between categories (i.e. over segmenting). Figure 3 shows the precision and recall measures for some representative systems based on CRF and MaxEnt respectively. With the CRF-based approach, the prompt features, the duration feature and the lexical features improve the segmentation performance. All feature sets here outperform the rule-based expert system in Table 1.

An interesting observation is that even without the lexical features, the CRF-based system still achieves an AED-ACC measure of 90.0%, which is similar to the best performance. This suggests that one can deploy a powerful segmenter without requiring an ASR system for the target language. Given the technical and financial challenges of deploying a full large-vocabulary ASR system, having a probabilistic trainable segmenter that does not depend on ASR brings significant practical benefits.

### 6. Conclusion and Discussion

We study the problem of high-level segmentation for customer call analysis using only call-center-side audio recording. We introduce a probabilistic trainable system based on a suite of features from speech activity detection, non-speech event analysis and speech recognition. We investigate two discriminative approaches, MaxEnt and CRF, and demonstrate that using a CRF-based method, the proposed system effectively uses the various features and outperforms a rule-based expert system. The proposed system can be rapidly deployed in a call center with a small number of annotated calls.
References


