Wireless Pers Commun (2014) 79:1959–1973
DOI 10.1007/s11277-014-1967-9

Typed VoIP Silence Prediction for Smartphone Energy Savings

Conner Kasten · Gang Zhou

Published online: 19 August 2014

Abstract Previous research has shown that considerable energy savings may be made in Android phones by intelligently switching the WiFi radio from Constantly Awake Mode (CAM), to Power Save Mode (PSM) during periods of mutual silence in Voice over IP calls. This is because a WiFi radio in CAM consumes approximately one third of a smartphones battery life. Since a typical conversation can consist of up to 60 % silence, much of a conversation need not be transmitted, and thus wastes a considerable amount of battery power. This research shows that as much as 40 % of that energy can be saved by switching the radio to PSM during mutual silence periods. However, this technique relies heavily on accurately predicting the duration of such silence periods. This paper presents and compares a silence model which uses speaker identity to build a typology of silences to a naive un-typed strategy. This type model relies on recent research showing statistical difference between the durations of classes of silences based on speaker identity before and after a silence. Evaluation of this technique, however, demonstrates that such models do not show improvement over more naive strategies. Thus, a naive strategy is perfectly sufficient for accurate prediction of silence duration.

1 Introduction

The use of Voice over IP (VoIP) calling on smart phones is becoming increasingly common, with over 1 million downloads of several such popular apps on the Google Play app market [4]. As such applications become more common, it becomes increasingly necessary to develop ways in which they may be improved, whether this involves improved call quality, or reduced energy consumption, without negatively impacting other features of the calling solution.

C. Kasten · G. Zhou (✉)
College of William & Mary, Williamsburg, VA, USA
e-mail: gzhou@cs.wm.edu

C. Kasten
e-mail: cgkasten@email.wm.edu

The use of WiFi during a VoIP call incurs a higher cost than a similar call over a cellular data network [14]. As such, it makes an ideal target for research attempting to reduce battery consumption during such calls, especially since another major drain, the microphone [13], cannot be switched off or duty cycled while maintaining call quality. These modifications to WiFi behavior must be made carefully, however. A quick, obvious solution to such a problem would be to make use of the Power Save Mode (PSM) which uses as little as one twentieth of the power of the standard Constantly Active Mode (CAM) [11]. This approach must be used judiciously, however, as PSM introduces significant latency (up to 300 ms [10]), which will negatively affect call quality. Because of this sensitivity to high latency, PSM may only be applied during periods when the absence or delay of information will be unnoticeable to users.

One obvious situation in which information transfer is unnecessary is during silent periods. Since up to 60 % of the average conversation is silence [3], there are considerable energy gains to be made. For instance, we can predict silence periods, and put the WiFi radio in PSM for the duration of such periods. Previous research (resulting in a solution called SiFi [11]) has demonstrated that this approach is an effective way of reducing the amount of energy used by VoIP calls, saving up to 40 % of energy consumed by such a call.

Previous work has show that conversational silence may be classified into types with different functions and distributional patterns [6,17]. Among this work is a study of silences which classifies based on speaker identity [5], a trivially easy piece of information to gather during VoIP calls. Further, this work shows that the durations of different classes of silence in this typology show very different distributional patterns. This strongly suggests that modeling each class independently should improve the accuracy of duration predictions.

This paper proposes TySi, Typed Silence WiFi energy savings, a model which uses easily accessible information about the identity of speakers. TySi takes this notion of speaker identity classification and uses it to predict silence durations during VoIP calls. This model is then compared to a previously proposed naive model [11], to determine whether speaker identity is useful information when predicting conversational silence. Comparison of the performance of the two models demonstrates that even though speaker identity has a statistical effect on silence duration, this information does not improve duration prediction.

The remainder of this paper is as follows. First, related work will be discussed, and used to examine the particulars of saving VoIP energy. Next, the silence typing model will be explained and examined. Experimental data will evaluate the feasibility of the models presented here. The paper will conclude with discussion of further possible work on the typed silence model.

2 Related Work

This section will focus on various previous approaches to gaining energy savings during VoIP calling. First, the underlying notion of foregrounded data is discussed. Silence typology models related to the TySi model follow. This section also examines some more general approaches to energy saving by tight management of WiFi radio.

2.1 Foreground/Background Data

Several approaches [1,10] rely on the notion of forgreounded data. This is time-sensitive data for which delay will cause reduced quality of service. Self-tuning [1] provides developers an opportunity to mark applications which require low latency, while SAPSM [10] determines

which data should be foregrounded using experimentation and user interaction. In both cases, a modified kernel driver places the radio in PSM whenever this higher priority traffic is not present.

The above work relies on the same basic insight as this paper, that some data is of higher priority to users. However, the work presented there is of broader focus, and does not assume, as this paper does, that some data may be dropped without consequence (as silence), because it's mere absence conveys the same information.

2.2 Silence Typology

Significant work has been done in certain subfields of linguistics to create accurate models of silence, going back to highly influential authors such as Rubinstein [15]. Since silence may be used for various pragmatic purposes within discourse, it has been proposed that silence may typed based on its discourse use [6].

However, such a typology relies on complex discourse functions such as rhetorical use [6], societal roles [18] or discourse structure [15]. As these functions are quite dificult to predict accurately (or cheaply) in a computational setting, a simpler typology must be used instead.

TySi uses a model focused on the turn-based structure of conversation [16]. This model divides silences into three classes, pauses, gaps, and lapses. A pause is any silence that occurs within a single speaker's turn. That is, a pause may be to think of a word or for rhetorical effect, but it is clear that the speaker is not done talking. In contrast, a gap occurs between turns. Thus, a gap starts when one speaker finishes speaking, and ends when the next speaker begins. A lapse is not defined in such clear terms, however, as being effectively defined as an especially or unnecessarily long gap. Later work [5] (and references cited therein) eliminates the lapse from the model, resulting in a clearer classification based entirely on pre- and post-silence speaker identity.

A recent study of the pause and gap durations in four speech copora consisting of three different languages (Dutch, Swedish, and Scottish English) demonstrated that the distributions of durations for these two silence types are noticeably different [5]. Their study showed that pauses tend to be longer than gaps (with the median pause being 100 ms longer), and further that while pauses are negatively skewed, gaps are positively skewed.

2.3 VoIP Energy Savings

TySi is essentially an extension of SiFi [11], a VoIP call energy savings approach focused on conversational silences. In this work, energy is saved by setting the WiFi radio to PSM during periods of silence. These periods are detected in one of two ways. When possible, Voice Activity Detection (VAD) is used to determine when silence occurs, since this protocol uses special packets to indicate when a silence period begins. This packet can be detected quickly without further analysis. When VAD is not available, SiFi uses a lightweight threshold based algorithm which determines when the audio level falls below a manually configured level. Note that in a noisy environment, this solution may result in less energy saving, since the algorithm does not differentiate conversational silence (when both participants cease speaking) and actual silence (when the volume is below a certain threshold).

Once such a silence period is detected, a machine learning algorithm determines the length of the silence, based on the set of silences gathered during a dynamically determined training period. The radio is put into PSM for the predicted duration. If the conversation is still silent when the radio comes out of PSM, the program calculates the probable length of continued silence, and puts the radio back to sleep for that duration, until speech is detected.

Another energy saving solution, GreenCall [9], operates under very different set of principles. It examines the difference between the timestamp of the packet currently being played to the user and the time of the most recently buffered packet. If the time is long enough, GreenCall puts the radio in PSM until new packets must be received. This reduces the cost of speaking packets, complementing the approach of SiFi and TySi.

3 Silence Typology Model

The silence model used for TySi is a basic extension of the turn-based model presented in Sect. 2.2. It maintains the distinction between pauses and gaps. However, it follows most current work in doing away with the lapse distinction. Thus, TySi's most basic typology is defined in terms of speaker identity before and after silence periods. That is, a silence is a pause if and only if the speaker directly after a silence is the same speaker as a directly before the silence. Similarly, a silence is a gap if and only if the speakers before and after the silence are not the same.

However, the pause/gap distinction is not the only one made by the TySi typology. This is because the pause/gap classification relies on the knowledge of the speaker after the silence has concluded, information which is obviously not available at duration prediction time. Thus, there must be some further information to make the pause/gap distinction useful. To that end, TySi uses the identities of speakers, a feature which, to our knowledge, is not used in other silence typologies.

Speaker identity is used in tandem with the pause/gap distinction in the following way. Consider that for any silence, there are two speaker identities of concern. The first is the speaker who stopped speaking directly before the silence began. The second is the speaker who ends the silence by beginning to speak. The TySi typology is interested in this second speaker, because it is she who decides how long the silence will be by ending it.

Every silence can be classified based on these two features: whether it is a pause or gap, and which speaker ends the silence. Thus, in a two person conversation, there are four types of silence: p_1, a pause ended by speaker 1(S_1), p_2 a pause ended by S_2, g_1 a gap ended by S_1, and g_2, a gap ended by S_2. These four silence types are illustrated in Fig. 1.

When a silence begins, the only information available to the predictor is who spoke directly before the silence. A p_1 (S_1 both starts and ends the silence) silence is indistinguishable from a g_2 (S_1 starts and S_2 ends the silence) silence at the silence's beginning. When TySi performs silence duration prediction, it has learned from sets of post-facto labeled silence periods. These sets are P_1, the set of p_1 silences, G_1, the set of g_1 silences, etc. So when TySi encounters a silence preceded by speech from S_1, it knows the following silence should either belong to P_1 or G_2. Thus, it's very typical to talk about the union of P_1 and G_2, which is referred to as U_1. The other union set for a conversation, combining P_2 and G_1, is U_2. Instead of the ending speaker (which is unknown) begin represented in the notation, the speaker preceding the silence is indicated in the union set.

S_1 is speaking	S_1's pause (P_1)	S_1 is speaking	S_2's gap (G_2)	S_2 is speaking	S_2's pause (P_2)	S_2 is speaking	S_1's gap (G_1)	S_1 is speaking

Fig. 1 This diagram illustrates the typing of silence by turn position. Assume that the conversation proceeds from left to right. *Shading* indicates that one speaker is talking

Obviously, it would be ideal to be consistent between the P, G, and U notations, but the nature of the union set makes this impossible. Note that all of this assumes that only two speakers are participating in the conversation. If there were three speakers, the U_1 set may also include G_3. If P and G sets followed the same notation as U (with the subscripted number indicating the speaker before the silence), then this U_1 set would contain P_1, G_1 (containing silences of gaps preceded by S_1 and followed by S_2, and a second G_1 (containing silences followed by S_3. Although the current work only considers two speaker conversations, the notation is selected so that an arbitrary number of speakers may be added to conversations while maintaining the necessary distinctions.

3.1 Silence Duration Prediction

Now that the basic typology has been examined, it remains to be demonstrated how this typology can be used to predict silence durations. In most ways, the techniques presented here are a direct descendant of the concepts used for SiFi [11]. The key differences lie largely in the selection and application of training sets, and will be emphasized at the appropriate points.

As discussed in Sect. 2.3, SiFi uses a machine learning algorithm trained on a series of conversational silences to predict a silence's duration at its onset. Specifically, an empirical cumulative distribution function (ECDF) is build to provide probabilities related to silence duration. A cumulative distribution function is a probability function for continuous distributions which takes as its input a particular sample from its target distribution and returns the probability that a sample be less than or equal to the input. Thus, for a uniform distribution from zero to ten, the CDF of five is one half. Here, the ECDF is build by a set of silences, and using them as the sample space to generate the function.

SiFi uses Kullback–Leibler divergence to determine when a sufficient number of silences have been collected by detecting when the ECDF converges. K–L divergence is a statistical measure of information divergence between two probability distributions. That is, it measures how much information is lost by approximation of a probability distribution. Here, it is used in sequence to see how much information can be gained by increasing the size of a model. When it's determined that adding more silences to a model will not increase information gain, SiFi can stop training because adding additional silences should not improve the approximation.

SiFi and TySi use the complement of the ECDF (that is, the probability that a sample will be greater than the input). That is, given that a silence has already lasted x, the algorithm returns a predicted duration δ, such that $\text{ECDF}(\delta|x) = \beta$. Here, β is a pre-defined parameter which represents the probability of the silence lasting δ given x. Thus, if $\beta = 0.7$, δ would be selected such that 70 % of silences are longer than δ. The Wifi radio can then be put in PSM for the duration of that predicted silence. When the predicted period ends, if the silence continues, a second prediction can be made, replacing x with $x + \delta$. This can be repeated until the end is detected or the ECDF predicts that the maximum duration has been reached [that is, when $x = \delta$ for ECDF $(\delta|x) = \beta$].

TySi differs from SiFi in that instead of building a single silence training set for prediction for all silences, TySi builds two per speaker, one for P silences and one for Gs. Then, a U set is built for each speaker, as discussed in the previous section. ECDFs are then generated for each U. TySi thus takes advantage of the knowledge of preceding speaker identity, by ruling out half of the silences that would be used by a SiFi predictor.

This approach as some additional advantages over the more naive SiFi approach. One of the most significant is the need for reatraining. If speaker identity does indeed have an effect

on silence durations, then SiFi must do retraining for every combination of speakers it deals with. That is, in such a situation SiFi must have separate predictors for conversations between S_1 and S_2, S_1 and S_3, and so on. However, once TySi has been trained for a particular speaker, all that is required to predict for a new speaker is to combine the pre-existing predictors for the speakers to generate new U sets (assuming both participants use TySi when making VoIP calls). Thus, training sessions should be far more limited for TySi (one per speaker) than for SiFi (one per possible combination of speakers). As long as a means for communicating P and G data for each speaker is included in an implementation, a very large amount of training time should be avoided without a significant reduction in prediction quality.

4 Model Evaluation

This section presents a comparison of the predictive capabilities of the TySi and SiFi models. This was achieved by creating a small corpus of VoIP calls and extracting silence durations to provide a trace set of real conversational silences. A larger corpus (as used in [5]) could not be used for this task because speaker identity between calls was needed to test the assumptions made by the model. The next sections outline the methodology for gathering the relavent data, and a presentation of the results garnered from statistical analysis of the traces recorded as well as trace-based simulation of the TySi and SiFi models to compare predictive power.

4.1 Trace Generation

Traces were generated from a set of ten 15-min VoIP calls made using Mumble [8], an open-source tool for VoIP selected for its attention to low latency, built-in use of VAD, and ability to provide per-user multi-channel audio recordings (making it trivial to determine which user is speaking at any point in a recorded conversation). The final result of a single recording is two uncompressed .wav audio files. Each audio file contains the recording of a single speaker. Further, the audio-files are time-synched, so that they are the same duration and begin and end recording at the same time. Recordings were made using laptops in quiet, but not soundproof rooms, using Microsoft 3000LX headsets.

Participants were students at The College of William & Mary. Further, all participants were members of the Science Fiction and Fantasy Club. This selection was done to ensure that all participants were familiar with each other and would be able to carry on two (relatively natural) 15 min conversations. No topic was specified, participants were told the could discuss anything they wanted, and the contents would be kept confidential. All participants were native speakers of English.

Each participant was paired at random with two others for two 15 min conversations. Thus, participant A (S_A) would first have a conversation with participant B, then later participant C. Letter codings are used to preserve speaker identity between conversations while protecting confidentiality. Figure 2 shows the participants' conversation pairings.

After all conversations had been recorded, acoustic analysis was performed using Praat [2], a popular toolkit for speech analysis. This analysis was broken down into a series of steps. First, intensity analysis was used to identify silence periods in the recording. This effectively transformed the audio recording into a textfile trace of time segments where acoustic intensity (volume) was below 60dB. Praat scripts available at the Speech Corpus Toolkit for Praat [7] were used to perform this task. As Mumble already performs VAD during the recording process, most silence periods were recorded at 0dB, making this process much

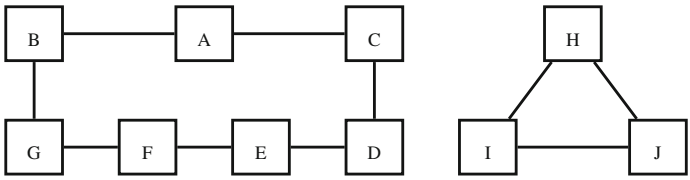

Fig. 2 Nodes represent speakers, edges represent a conversation was recorded between the linked speakers

quicker and more accurate. Visual inspections confirmed that almost no silence was labeled as non-silence, or vice versa.

The textfile output of the Praat script was then converted to a tabular format using a Python script which reduced each silence to a duration, preceding speaker, and following speaker. The script preserved the ordering of the silences for use in simulations and to evaluate whether there was some lag or autocorrelation between silences in the series. This final format was then imported into R [12], where the remainder of the analysis was performed.

4.2 Dataset Description

The final dataset consists of ten sets of labeled silence durations. The labels include the identity of the speaker before and after each silence. The number of silences in each subset ranges from 189 to 350, with a median of 226.5, and a total of 2,364 silence periods. The durations of silence periods range from 0.004 to 4.188 s, with a mean of 0.321 s and a median of 0.116 s. The mean of pauses was 0.247 s, with a range from 0.036 to 4.012 s, and the mean of gaps was 0.462 s, with a range from 0.004 to 4.188 s.

Interestingly, this conflicts with the findings referenced earlier [5], which finds pauses to have a higher mean duration than gaps. However, the authors cite several other studies on silence durations, some of which agree with this set, showing gaps to be longer than pauses. This suggests that the duration of pauses and gaps varies largely from corpus to corpus. If the cause of this variation is speaker-identity related, this could bode very well for the performance of TySi modeling when compared to SiFi across a large group of speakers.

In evaluating the health of the data, it was observed that one dataset was unusually leptokurtic. Conversation C_{HJ} (the conversation between S_H and S_J) has a kurtosis of 27.7, where the average kurtosis for the set is 12.4. While this suggests that something unusual happened during this conversation, the silence duration models should be able to handle such unusual behavior. As such, it is left in the dataset, and not thrown out as an outlier. However, it is worth paying close attention to this model in subsequent examinations to see how well TySi and SiFi handle unusually distributed silence durations.

4.3 Analysis of Variance

The first step in confirming that the TySi model is useful for predicting silence durations is ensuring that the underlying assumptions of the model are correct. Previous work demonstrated that the pause/gap distinction has real and measureable effects on silence durations [5]. However, this and other work does not attempt to assure that speaker identity affects silence duration in a similar way (this is largely because other factors are assumed to be more important). As such, an important first step is to confirm that speaker identity impacts silence duration in a meaningful way.

As discussed in Sect. 3, TySi assumes that the identity of the speaker who begins talking directly after a silence period has a significant impact on the duration of the silence. An analysis of variance is used to determine whether or not this is the case. For this test, the dataset was divided into two subsets, one consisting of all gaps, and the other consisting of all pauses in the dataset. An analysis of variance was performed on each set to determine if following speaker identity had a significant effect on silence duration. The p values for both sets were approximately 0.0002, falling well below the significance threshold of 0.05. This suggests that following speaker identity does indeed have a measureable effect on silence duration. Thus, we are able to affirm that TySi's model relies on linguistically real features when predicting silence durations.

4.4 Cross Validation

The next step of evaluation is to determine whether or not a silence duration model using the TySi design accurately predicts silences. Ten-fold cross validation is used to confirm that a well-trained TySi model accurately predicts silences. In such a test, a target dataset is split into two sets, with one set consisting of 90 % of the data. This is the training set, on which the model will be trained. The remaining 10 % is a test set on which the model will be tested. This process is repeated ten times (thus 'ten-fold' cross validation), which each repitition using a different test set. Note that here the test set is a randomly selected, non-contiguous set of silence periods, and that there was no overlap between the ten test sets.

This process was repeated for each conversation, so the TySi model was built only out of silence durations from one conversation, and tested only on silence conversations from that same conversation. The R^2 coefficient of determination was used to determine how well the TySi models built during cross-validation were able to predict the test sets. This coefficient is used to compare the variability of estimation errors with that of original values being estimated. As it has varying definitions depending on the use case, the formula used to estimate R^2 in this case is presented in Fig. 3.

In this formulation SSE is the sum of squared errors, composed of the squared differences between each real value and the prediction TySi produced for that value. SST is the total sum of squares, composed of the squared differences between each real value and the mean of the dataset being tested. The resultant R^2 value is a decimal between zero and one. As the estimations made by TySi improve (trend towards the real value), the R^2 value trends toward one. Thus, R^2 values close to one indicate the model makes accurate estimations.

Table 1 presents the results of the ten-fold cross validation. In addition to cross validating the TySi model, The "Full" two dimensional, four category silence model presented in Sect. 3 was examined. Of course, this model is not useful for the practical application of VoIP energy saving as TySi is. Nevertheless, TySi's model is built directly on this underlying notion, and it would be comforting to confirm that it does in fact work as well or better.

Fig. 3 The formula for R^2 values

$$SSE = \sum_i (y_i - f_i)^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

Table 1 R^2 results of tenfold cross-validation

Conversation	TySi	Full
C_{AB}	0.9966	0.9859
C_{AC}	0.8965	0.9873
C_{BG}	0.8632	0.9803
C_{CD}	0.6454	0.9913
C_{DE}	0.9912	0.9959
C_{EF}	0.9997	0.8321
C_{FG}	0.9916	0.8576
C_{HI}	0.4505	0.5502
C_{HJ}	0.6519	0.9665
C_{IJ}	0.4294	0.9136
Mean	0.7916	0.9061

This data is something of a mixed bag. Although most of the "Full" model predictions are quite accurate, four of the conversations seem to have presented a significant challenge to TySi. In particular, C_{HI} not only presented a challenge to TySi's predictive abilities, but to the full model as well. In fact, all of the conversations in the $H-I-J$ ring presented some problems for TySi. This ring was recorded a week after the larger $A-B-C-D-E-F-G$ ring of conversations due to the availability of participants for testing. However, the recordings were made in the same facilities using the same equipment. Further, other than C_{HJ}, no unusual statistical properties were discovered related to these conversations or C_{CD} (the other conversation with a TySi R^2 value <0.8).

Although these conversations do bring up some significant concerns about the robustness of the TySi model, relatively high mean R^2 values suggest that all hope is not lost. The TySi model is still capable of predicting silence durations with very high accuracy in a majority of conversations. Further, this test did not attempt to compare TySi's predictive capability with SiFi's, and does not take into account the advantages that TySi's modeling has of SiFi's, such as the ability to exchange user-specific models to avoid training on each conversation. Thus, while the results of the cross validation are not ideal, they are accepted as "good enough" to justify comparing the proposed model to SiFi.

4.5 Simulations

Because of the randomized nature of cross validation, it is not suited for comparison of two modeling techniques. Similarly, two silence predictors cannot be run on the same VoIP conversation in real time, ruling out a comparison of implementations. Trace-based simulation offers a solution which allows direct comparison of models on identical data with access to the same set of information.

In the following experiments, simulations are performed on the traces of silence periods derived from recording. First, the two predictive models to be used for each conversation are generated using the TySi and SiFi techniques (here, the "full" model is left untested, as it requires information that would not be accessible during the real-time experiment being simulated). Next, prediction of the duration of each silence is performed in the order it appeared, as outlined in Sect. 3.1.

The results of the following simulations allow us to directly compare how well TySi and SiFi models are able to predict silence duration. Once again, results will be presented

using R^2 values. Thus, a simple comparison of which model received a higher R^2 for each conversation indicates which model produces the better predictions for that conversation.

4.5.1 Within Conversations Simulation

This first simulation examines the models' abilities to predict silence durations when only a single conversation is available. Such a situation will of course occur the first time a user uses a VoIP client which uses either TySi or SiFi. After this first conversation, TySi would not need to use such a strategy again, as it can use models from a previous conversation.

In this simulation, TySi and SiFi models were built for each conversation, using the first half of silences present (where silences are ordered temporally as they occurred in the original recording). Kullback–Liebler divergence is calculated as described in Sect. 3.1 to confirm that this division produces models based on sufficient data to be useful.

Following this, predictions are made on the second half of the silence set, and R^2 values are calculated for the predictions made by TySi and SiFi respectively. The α and β parameters were set at 0.05 and 0.3, values empirically derived both by the authors of SiFi [11], and independent empirical evaluation showing these to be optimal settings for this dataset.

Table 2 presents the results of this simulation. Both models perform quite well, with each outperforming the other in some instances. Overall, SiFi pulls slightly ahead, by 0.02. However, it seems unlikely that such a small difference would be significant.

C_{HJ} is not presented in the table, as R^2 values for this conversation for either model were extremely poor. In fact, both R^2 values were negative, which indicates that the models performed worse than if the prediction of every silence duration were simply the mean duration of the training set. This suggests that the highly peaked distribution of this set was indeed problematic for predictive models like TySi. SiFi fairs no better, however, suggesting that whatever anomaly caused this conversation will negatively impact any attempt to model silence here.

Of the other conversations which presented some problems during cross validation, it appears that only C_{IJ} still is a challenge to predict. There do not appear to be any particular statistical properties which set this subset as a whole apart from the others. Further examination discovered that the kurtosis of the training set was almost twice that of the test set (10.5 vs. 5.9).

This suggests that it is possible for the distributional properties of conversational silence to change over the course of a conversation. Further examination showed only C_{IJ} in this

Table 2 R^2 values for models tested in the within conversations experiment

Conversation	SiFi	TySi
C_{AB}	0.9377	0.9488
C_{AC}	0.9877	0.9542
C_{BG}	0.9935	0.8572
C_{CD}	0.9993	0.9999
C_{DE}	0.8091	0.9493
C_{EF}	0.9983	0.9594
C_{FG}	0.9820	0.9929
C_{HI}	0.9466	0.8444
C_{IJ}	0.5983	0.5849
Mean	0.9169	0.8990

data-set, with all others having roughly similar distributions between test and training sets for the other nine conversations. However, for conversations where this effect does occur, training only at the beginning of a conversation is deeply insufficient. This suggests that different techniques should be used when selecting training sets for such models.

4.5.2 Between Conversations Simulation

The second simulation examines how TySi performs when trained on data from a different conversation. This should be a particular strength of TySi. When two speakers have a conversation on a TySi system for the first time, TySi can use previous models for those speakers and integrate them to generate new U sets unique to that pair. In contrast, SiFi must use models that have at least one different speaker. If speaker identity really matters to when predicting silence duration, as suggested by the previous analysis of variance, an experiment where previous conversations' information data are used should should TySi performing remarkably better than SiFi.

In this series of simulations, models are built using "previous" conversations. Recall that in the data traces used, each speaker participated in two conversations. Thus, S_A participated both in C_{AB} and C_{AC}. For this simulation, models are built using the entire trace of a separate conversation. Thus, when simulating C_{AB}, models are built using C_{AC} and C_{BG}. TySi first generates P and G sets for each conversation, then combining them to create new U sets for S_A and S_B, as outlined in the Models section.

Since SiFi cannot intelligently combine these models, the simulation of the conversation using SiFi silence duration prediction assumes that both speakers' devices are running SiFi, using models built from their respective "previous" conversation. This results in two sets of predictions, one for each speaker. In our example above, one set of predictions is based on a SiFi model trained on C_{AC}, and one uses a model trained on C_{BG}. This is how models trained on previous conversations are handled in the SiFi paradigm.

The results in Table 3 are roughly comparable to in Table 2. However, mean values show some across the board improvement over the within conversations experiment. This is somewhat surprising, as SiFi research showed a mild reduction in performance when models where applied between conversations. This may be the result of a simple increase in data available in the training set,. However, the K–L convergence used to determine when to stop training showed that all of the tests in the within conversations experiment had sufficient

Table 3 R^2 value results of the between conversations simulation

Conversation	SiFi 1	SiFi 2	TySi
C_{AB}	0.9899	0.8022	0.9791
C_{AC}	0.9516	0.9340	0.9634
C_{BG}	0.9997	0.9873	0.9973
C_{CD}	0.9182	0.9530	0.9697
C_{DE}	0.8454	0.9924	0.9884
C_{EF}	0.9843	0.9996	0.9703
C_{EG}	0.9501	0.9796	0.9962
C_{HI}	0.9674	0.9176	0.7521
C_{IJ}	0.7235	0.8045	0.8543
Mean	0.9145	0.9300	0.9412

Table 4 R^2 values for the mixed-training simulation

Conversation	SiFi	TySi
C_{AB}	0.9991	0.9841
C_{AC}	0.8944	0.9177
C_{BG}	0.9690	0.9378
C_{CD}	0.9993	0.9999
C_{DE}	0.9953	0.8637
C_{EF}	0.9998	0.9998
C_{FG}	0.9690	0.8534
C_{HI}	0.9889	0.9888
C_{IJ}	0.8292	0.8505
Mean	0.9604	0.9328

data to predict reliably, and further that more data of the same kind would not improve performance.

This suggests that some different information is being provided later in the conversation that would not be accounted for by K–L divergence. This is consistent with the observations about C_{IJ}'s kurtosis changing over the course of the experiment. Although tests to examine time-correlation did not suggest a significant pattern in duration change over time, it seems quite possible that the results here are the consequence of full conversations being the training material, rather than partial conversations.

4.6 Mixed-Training Simulation

The suspicion that training on a full conversation disproportionately improves model performance is tested in a third simulation. In this experiment, models are trained on full conversations once again, but they are full conversations with no speakers in common with the test conversation. This is in contrast to the previous two conversations, in which training data always shared at least one speaker with test data.

Each conversation was tested using models built from C_{FG}, except for the three conversation with at least one of these speakers participating. These conversations were tested with models built using C_{AC}. TySi models arbitrarily assign the F models to the earlier alphabetical participant, and G to the other (with an equivalent procedure carried out for C_{AC} modeled conversation). Test results are given in Table 4, in the same format as previous simulations.

The results of this simulation make it clear that models built using a full conversation are indeed the deciding fact which improves the performance of the between conversations simulation over the within conversations simulation. Even so, the improvement is very slight for eight of the nine conversations, with only C_{IJ} seeing marked improvements when a full conversation is used. That is, for most conversations, a simple SiFi model build on a set of silences determined to be sufficient by K–L convergence is good enough.

5 Discussion

The simulations given in the previous section lead to some surprising conclusions. First, although statistical testing indicates that speaker identity is significantly correlated with silence duration, incorporating some of this information into a predictive model does not appear to significantly improve results. This is likely because although some of this infor-

mation can be accessed at the time of prediction, the critical question of who will end the silence is not available until it is over. Thus, the model must also use the statisically significant pause/gap distinction. However, it seems that without the identity of the ending speaker, these distinctions are not strong enough to make a significant increase in predictive power. Thus, TySi's model does show significant improvement over SiFi.

The second, somewhat more unexpected result is that K–L divergence cannot always correctly determine when the model training has maximized information gain. Comparison between the within conversations simulation and the between conversations simulations suggest that in some small subset of cases (one of ten total conversations), a full conversation is needed to build accurate models. This suggests that some, but not all, conversations do see changes in duration patterns over time. However, it appears that this dataset is not large enough to uncover such patterns, as investigation into this possibility yielded no significant results.

A third observation to be made from the unfortunate removal of C_{HJ} from the dataset suggests that some conversations may simply not have silence durations which are well modeled by this approach. No experiment aside from cross-validation was able to perform well on this conversation. It is possible that multiple conversations between these partners may result in a model able to predict their silence patterns, and more data is required to examine this. However, K–L divergence indicated that sufficient data was available in the models to make good predictions, suggesting that some other force, possibly pragmatic or sociological in nature, is affecting the silence periods with these speakers.

6 Conclusion

In conclusion, this research has attempted to determine whether speaker identity, coupled with the pause/gap distinction, can be used to improve the prediction of silence durations during VoIP calls. Such improvements could lead to higher energy savings and/or better call quality when such techniques are implemented to save energy on mobile devices performing VoIP calls over wireless networks. However, the results show that such information does not provide a significant improvement to such predictions. Nevertheless, the experiments exposed potential for a different avenue of improving directions. It seems that some subset of conversations are not well modeled by existing techniques due to changes in silence models over the course of a conversation. The results show that a full conversation is needed as training material to effectively model this subset.

7 Future Work

The determination that speaker identity does not appear to provide useful information in predicting silence duration actually opens new opportunities for research. As mentioned earlier, large corpora of phone-call recording data do not include speaker identity between calls. This makes them unhelpful when testing TySi, but perfectly acceptable for testing other models. Thus, future testing can make use of available corpora (such as those referenced in [5]), which are significantly larger than the corpus built for this project.

These larger corpora can then be used to test new learning and retraining techniques to resolve the problems observed in predicting C_{IJ}. Since the standard technique for building predictor models in SiFi is to use the first n silences in a conversation (until K–L divergence tests indicate sufficient data has been gathered), this conversation and others like it would not

be easily predictable for the SiFi model. Thus, one of two solutions must be implemented to handle such cases. The first is to find some new training technique that guarantees that the models will handle such cases. The second is to determine some type of retraining detection technique to recover when such a conversation is encountered. The availability of large corpora to test these techniques with makes evaluation and comparison of a variety of such techniques quite easy.

References

1. Anand, M., Nightingale, E. B., & Flinn, J. (2005). Self-tuning wireless network power management. *Wireless Networks*, *11*(4), 451–469.
2. Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.
3. Drago, P., Molinari, A., & Vagliani, F. (1978). Digital dynamic speech detectors. *Communications, IEEE Transactions on*, *26*(1), 140–145.
4. Google Play. (2013). http://play.google.com/
5. Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555–568.
6. Kurzon, D. (2007). Towards a typology of silence. *Journal of Pragmatics*, *39*(10), 1673–1688.
7. Lennes, M. (2011). Spect-The speech corpus toolkit for praat. http://www.helsinki.fi/lennes/praat-scripts/
8. Mumble. (2013). http://mumble.sourceforge.net/
9. Namboodiri, V., & Gao, L. (2008). Towards energy efficient voip over wireless lans. In: Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing, ACM, pp. 169–178.
10. Pyles, A. J., Qi, X., Zhou, G., Keally, M., & Liu, X. (2012). SAPSM: Smart adaptive 802.11 PSM for smartphones. In: Proceedings of the 2012 ACM conference on ubiquitous computing, pp. 11–20.
11. Pyles, A. J., Ren, Z., Zhou, G., & Liu, X. (2011). Sifi: Exploiting voip silence for wifi energy savings insmart phones. In: Proceedings of the 13th international conference on Ubiquitous computing, ACM, pp. 325–334.
12. Core, R., & Team (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
13. Rahman, M. M., Ali, A. A., Plarre, K., al'Absi, M., Ertin, & E., Kumar, S. (2011). mconverse: Inferring conversation episodes from respiratory measurements collected in the field. In: Proceedings of the 2nd conference on wireless health, ACM, p. 10.
14. Rahmati A., & Zhong, L. (2007). Context-for-wireless: Context-sensitive energy-efficient wireless data transfer. In: Proceedings of the 5th international conference on Mobile systems, applications and services, ACM, pp. 165–178.
15. Rubinstein, A. (1998). *Economics and language*. The Schwartz Lecture, Northwestern University.
16. Sacks, H., Schegloff, E. A., Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, pp. 696–735.
17. Wilson, T. P., & Zimmerman, D. H. (1986). The structure of silence between turns in two-party conversation. *Discourse Processes*, *9*(4), 375–390.
18. Zimmerman, D. H., West, C., et al. (1975). Sex roles, interruptions and silences in conversation. *Language and Sex: Difference and Dominance*, *105*, 129.

Conner Kasten received a Bachelor's degree in Linguistics from the College of William & Mary. He is currently a Master's student in the Computer Science department there.

Gang Zhou is an Associate Professor in the Computer Science Department at the College of William and Mary. He received his Ph.D. degree from the University of Virginia in 2007 under Professor John A. Stankovic. He has published over 50 academic papers in the areas of sensor networks, smartphones and ubiquitous computing, and low power wireless communication and networking. The total citations of his papers are more than 3,500 according to Google Scholar, among which three of them have been transferred into patents and the MobiSys'04 paper has been cited more than 690 times. Ten of his papers have each attracted more than 100 citations since 2004. He served as technical program vice chair, session chair, member, and doctor colloquium panelist for 56 academic conferences, including ACM SenSys, IEEE INFOCOM, IEEE RTSS, and IEEE MASS. He has served as NSF, NIH, and GENI proposal review panelists multiple times since 2008. He received an award for his outstanding service to the IEEE Instrumentation and Measurement Society in 2008. He also won the Best Paper Award of IEEE ICNP 2010, which was given to only one paper from among the 170 papers submitted (acceptance rate: 18 %). He received NSF CAREER Award in 2013.