

# Interpersonal Relationship Labels for the CALLHOME Corpus

Denys Katerenchuk<sup>1</sup>, David Guy Brizan<sup>1,2</sup>, Andrew Rosenberg<sup>3</sup>

<sup>1</sup>CUNY Graduate Center

365 Fifth Avenue, New York, NY 10016

<sup>2</sup>University of San Francisco

101 Howard Street, San Francisco, CA 94105

<sup>3</sup>IBM Research AI

1099 Kitchawan Rd., Yorktown Heights, NY, 10598

<sup>1</sup>dkaterenchuk@gradcenter.cuny.edu, <sup>2</sup>dgbrian@usfca.edu, <sup>3</sup>amrosenb@us.ibm.com

## Abstract

The way we speak to our friends, colleagues, or partners is different in both the explicit context, what we say, and the implicit, how we say it. Understanding these differences is important because it provides additional information that can be used in natural language processing tasks. For example, knowing the relationship between interlocutors can help to narrow the range of topics and improve automatic speech recognition system results. Unfortunately, the lack of corpora makes exploration of this problem intractable. In this work, we release a set of interpersonal relationship labels between conversation participants for the CALLHOME English corpus. We make the labels freely available for download on our website and hope that this effort can further boost research in this area.

**Keywords:** corpus annotation, interlocutor relationship labels, CALLHOME corpus

## 1. Introduction

Communication is one of the most invaluable tools humans have. It enables us to understand each other, share ideas, coordinate work, and build relationships. Through speech, we carry explicit information in terms of words, as well as implicit information that is usually expressed by an acoustic signal of the human voice. For example, when you are on the phone with a friend, it is often clear when the person is distressed, even when they are trying to conceal or deny this fact. Word choice and speech intonation are formed on the conscious and unconscious levels. Unlike the conscious level, we have little control over the unconscious. Understanding the relation between explicit and implicit information during conversations can provide additional information about the speakers.

We have all been in a park or a restaurant and overheard or observed strangers interacting with each other. Have you ever tried to guess the relationships between them? Were they two old friends, cousins, or maybe a couple? On what distinct characteristics did you base your assumption? Was it the body language, voice, or word choice? It may seem like an easy task for humans to identify the relationships. However, this is a very difficult task for computers.

People are interacting with cellphones, smart TVs, and computers on a daily basis using voice-based interfaces. However, these interactions can be harmed by misunderstandings. One reason for the occurrence of the misunderstandings is that these systems rely on automated speech recognition (ASR) systems, which, despite showing strong improvements in performance, are far from perfect. One way to improve an ASR system's performance is to take into account not only spoken words, but also speaker and domain information. For example, information about speaker's gender, age, or voice characteristics can be used to reduce recognition word error rate (Saon et al., 2013;

Serizel and Giuliani, 2017). The knowledge of a discussion topic, in a similar fashion, can aid the accuracy (Chen et al., 2015). If conversations involve multiple speakers, knowing the relationships among the participants may be beneficial because many times the topics and voice differ when we converse with colleagues or family members. Unfortunately, there are very few resources available in this area. As noted by (Kendall, 2011), many existing corpora lack the labels researchers need to investigate the effects that interlocutor relationships play in language change.

The main focus of this paper is to provide a set of labels to boost research of language and its differences between family members and friends. The results of such research can be used in improving multiple NLP areas. We release a set of annotated labels for the well-known CALLHOME English corpus of phone dialogues (Canavan et al., 1997). The labels are available for download at [https://github.com/dkaterenchuk/callhome\\_labels](https://github.com/dkaterenchuk/callhome_labels).

## 2. Related Work

In recent years, a great deal of notable research has been done on studying implicit information from speech conversations and written dialogues. The early work in this domain by Stirman and Pennebaker (2001) has shown that there is a correlation between word choices and the mental states of the authors. Their work analyzed poetry documents to identify suicidal writers. They found that these authors tend to use more words that are related to themselves rather than to others. Another paper, from authors in the same research group, showed that it is possible to identify the level of romantic interest during a speed dating session and also to predict the likelihood of a long term relationship (Ireland et al., 2011). This work is based on the analysis of word choice and language style, known as linguistic style matching.

Speech contains a rich source of implicit information and a lot of work has been done to study its communication. For example, Rao et al. (2012) and Han et al. (2014) (among many others) show that voice can carry information about emotions. Polzehl et al. (2010) proposed a method for predicting a speaker’s personality traits. This information about speakers can be used to create personalized responses of conversational agents as described in the work by Siddique et al. (2017). In addition to speaker information, our voices carry data about intent and deception, as was shown in the work by Sanaullah and Gopalan (2013), Levitan et al. (2015b) and Mendels et al. (2017). The ways we converse with coworkers or partners is also unique. The study of Spanish phone conversations by Yella et al. (2014) shows that with the accuracy of 75%, it is possible to recognize if a conversation is between partners or family members. Previously, we studied a similar problem of identifying relationships between friends and relatives (Katerenchuk et al., 2014). The results confirm that the way we speak to our friends is different from conversations amongst family members.

These research efforts and their outcomes were made possible by the data availability. For example, the release of the SpeedDate corpus (Ranganath et al., 2009) made working on investigation of romantic interactions possible to researchers. Similarly, Maekawa et al. (2000) and Campbell (2002) collected spontaneous speech of Japanese speakers that lead to improvements including phoneme recognition (Fourtassi et al., 2014), domain adaptation (Asami et al., 2017), etc. The most similar dataset to ours is the CallNotes corpus (Carrascal et al., 2012). This is a set of phone conversations designed for speech analysis. The main difference is that the data is collected from native Spanish speakers. The dataset of our choice is based on phone conversations too, but the conversations are collected from native English speakers. Through this work, we hope to encourage research in understating vocal and textual differences between conversation participants.

### 3. Corpus Description

This section provides an overview of our decision to work with the CALLHOME English corpus, the description of the dataset, and the process for interpersonal relation label creation.

#### 3.1. Data Requirements

Data collection is often an expensive and time-consuming process. For this reason, we decide to look at available English dialogue corpora. The CALLHOME English corpus (Canavan et al., 1997) is a well known resource of English phone conversations. The main advantage of this dataset is that it complies with our requirements:

- The dialogs are in English.
- The conversations are dyadic<sup>1</sup>.

---

<sup>1</sup>There is one case in which multiple speakers appear on the call. This is noted in our labels.

- The speech is spontaneous.
- The participants are free to discuss any topic.

These requirements are important for the following reasons: 1) the annotators speak English and can perform labeling task reliably, 2) the real-world conversations that are not enforced by a specific topic provide natural research environment for the future analysis. For these reasons, we find the CALLHOME English corpus is a great choice of data for our annotation task.

#### 3.2. Data Description

The CALLHOME English corpus was developed by the Linguistic Data Consortium (LDC) and contains 120 unscripted phone conversations between native English speakers. The speakers are representatives of various demographic groups. The conversation participants were aware of the recordings; however, the conversations were on any topic of their choice and did not have any additional constraints. All phone calls were placed from North America to friends or family members who largely live outside of the USA and Canada. Each phone conversation is around 30 minutes in length for a total of 56.7 hours of audio. The conversations are divided into train (80 conversations), development (20) and test (20) sets.

The CALLHOME English corpus also provides transcripts. The transcripts cover a continuous 5 or 10 minute segment taken from a recorded conversation. The total time of transcribed audio is 18.3 hours. The transcribers were given a set of instructions that limit the transcribed segment to the middle of the conversation, preserve disfluencies, sounds, simultaneous speech and mispronunciations. Additional instructions and corpus descriptions appear in Canavan et al. (1997).

The corpus also provides anonymized speaker data. The information, presented in the corpus, describes speaker’s call ID, gender, age, years of education completed, state where the speaker grew up, and country or area code with first three digits of the dialed number. While the corpus supplies speaker information, it omits any data about interpersonal relationships between the speakers.

#### 3.3. Annotation

The annotations we provide were performed by a group from the Speech Lab @ Brooklyn College, CUNY (formerly of Queens College, CUNY). The annotators were asked to listen to the full conversations and refer to the transcripts, where available, to identify relationships between the call participants. The decision for each label is based on evidence from the conversation. The evidence could be a spoken or transcribed phrase such as “our parents” that signifies the speakers are siblings or a direct speech, such as “hello mom,” that shows that the conversation is between a parent and a child. Annotators described the relationship using any term they like. However, all annotations were entered into a shared document, which led to a relatively rapid convergence to a small set of labels.

Despite this, there are still some individual differences in the labels that are resolved after annotation is completed.

FRIEND	RELATIVE	
80	28	
FRIEND	SIBLING	PARENT-CHILD
80	15	13

Table 1: Label distribution

We find that most conversations are between friends – some of whom could be identified as work colleagues. We ultimately settled on two binary interpersonal relationships, FRIENDS and FAMILY, for the main label set. The line between these groups can be very thin since very close friends may feel like relatives and cousins or siblings may also be friends.

We were unable to find a finer-grained distinction of *types* of friends reliably across the whole corpus. As a result, the friendship sub-categories are not available and the conversation is labeled as “friend” in both cases. One conversation, numbered 5046, stands of an exemplar of the reasons why: the participant friends showed familiarity with each others’ families, the details of their homes and obligations to send presents in celebration of birthdays—a friendship of pleasure. Despite this, our annotators also determined that the relationship likely started with the pair having worked together—a friendship of utility. If we employ Aristotelean friendship categories, this relationship likely falls into at least two bins. We find many examples of this complex, multi-class friendship type in the CALLHOME corpus.

In the case of family members, in contrast to friends, we provide additional labels that further define the relationships. These additional labels consist of relationships such as mother, father, sister, brother, and cousin for each participant of the call, where they could be determined.

The annotation task is non-trivial in many cases. We are unable to provide labels for 12 conversations (10% of the corpus) because 1) the relationship cannot not be determined with confidence or, 2) in two instances, more than two speakers joined the conversation. These situations cause the interpersonal relationship between the speaking parties to change over the course of the conversation. An interesting quality of the CALLHOME data is that a small number of the conversations is between representatives of a religious group who refer to each other as “sisters,” when they are actually friends or colleagues. In these cases, the annotators have to find additional evidence of the relations and disregard these direct addresses.

In total, there are 108 annotated phone conversations. A summary of the data annotation can be found in the Table 1. The majority of instances, 80 out of 108, are labeled as FRIEND. The remaining 28 conversations are between family members and labeled as RELATIVE. The finer grained distinction between relative types is defined

by 15 instances of conversions between siblings and 13 between parents and children. This creates a highly unbalanced corpus. For this reason we provide the labels as a single set without a division for training, developing and testing subsets. We leave the normalization method or an appropriate use case of the data up to the user. The annotations of the CALLHOME English corpus are available at [https://github.com/dkaterenchuk/callhome\\_labels](https://github.com/dkaterenchuk/callhome_labels).

## 4. Data Analysis

We report our initial results on classifying interpersonal relationships that appeared in our previous work (Katerenchuk et al., 2014). During this initial exploration, we use a subset of the annotated data. The data consists of 56 phone conversations where 28 conversations are between friends and 28 are between relatives. Furthermore, we use 10-fold cross validation during the classification. In our experiments we use acoustic and textual data representations.

Our acoustic data representation pipeline is based on openSMILE, an open-source tool (Eyben et al., 2010). OpenSMILE provides a set of configuration files for acoustic feature extraction. We use the emotion.conf configuration from IS09 (Schuller et al., 2009). This configuration extracts 384 features that includes five low-level descriptors (LLDs) of acoustic features: 1) Zero crossing rate, 2) RMS Energy, 3) F0, 4) Harmonic-to-Noise Ratio, and 5-16) 12 MFCC coefficients. The change ( $\Delta$ ) of each of these LLDs is also calculated. This leads to a total of  $16 \cdot 2 = 32$  LLDs. Twelve functionals are then applied to these: 1) mean, 2) standard deviation, 3) skewness, 4) kurtosis, 5-8) value and relative position of minima and maxima, 9) range between minima and maxima, 10-12) linear regression coefficient, offset and MSE.

Textual representation is extracted from the transcripts. Since we wanted to investigate the relationships, we use a set of words proposed by Chung and Pennebaker (2007). In their work they show that function words, such as pronouns, articles and prepositions, are highly correlated with the speakers’ attributes. The counts for each of these is used as a representation. In addition, we use turn-taking information, interruptions, cuts off, delays in response, and other conversation related data.

The problem of identifying interpersonal relationships is cast as a classification task. The models are trained using both acoustic and textual data representations. We would like to point out that during our exploration the feature set is larger than the number of data points. In the real word system this set up is not ideal and might lead to overfitting. However, it should provide an intuition for further exploration. The goal is to investigate if any features contain predictive information and can identify the relationships between the speakers. The choice of our learning algorithms was limited to: 1) SMO, an SVM optimization algorithm, 2) J48, a decision tree algorithm, 3) Naive Bayes, and 4) BayesNet, a Bayesian Network

Classifier	Acoustic Feature	Text Features	Acoustic + Text	Acoustic Side A	Acoustic Side B	Acoustic Segment
SMO	42.9%	57.1%	39.3%	37.5%	48.2%	48.2%
J48	55.6%	57.1%	60.7%	60.7%	62.5%	57.1%
Naive Bayes	44.6%	60.7%	57.1%	50.0%	37.5%	50.0%
BayesNet	55.4%	55.4%	51.9%	55.6%	73.2%	51.8%

Table 2: Results

learning algorithm. In addition, we create experiments to analyze different settings of conversations and answer the following questions:

1. Can we identify relationships from a conversation?
2. Do we need to hear both sides of the conversation?
3. Is the whole conversation required to make a prediction?

The results of the experiments are shown in Table 2. From the table we can see that providing a full conversation, both acoustic and textual representations are indicative of the speaker relationships. However, the text based representation seems to have more information for this task, yielding 60.7% accuracy when using the Naive Bayes algorithm. Combining both representations achieved the same accuracy but with a different learning algorithm. From the analysis of the features, the MFCC-based acoustic signal is the most informative of the relationships. An interesting fact was discovered from transcript extracted text features. We found that conversations between friends are more egocentric and are reflected in higher frequencies of personal pronouns such as “my” and “I” (Table 3). In contrast, relatives appear to be more likely to discuss other people, which corresponds to a higher usage of third person pronouns. For more details on the feature importance we refer the reader to our previous work (Katerenchuk et al., 2014).

From the analysis of only one side of a conversation, we find that predictive results improve and produce the accuracy of 73.2%. This stronger result, however, comes with a caveat – only one speakers of the pair shows a strong predictive signal. In the case of current dataset, speakers receiving the call show higher predictive results. It is possible that this is a phenomenon of the distribution of data or that it can be attributed to the specifics of the callers; in this case that the speaker on side A places a call to a speaker on side B, who is likely located outside North America and who may share experiences which are more likely to be classified correctly.

Lastly, we explore the case where only a part of a conversation is available. From each audio, we extract a segment of 10 minutes from the middle of a conversation. We find that the accuracy increases in the majority of cases. This can be attributed to a number of possible causes including the fact that the speakers can be uncomfortable with being recorded and thus tend to be cautious at the start of the conversation,

Text Feature	Gain Value	Dir.
my_PER_(TOKEN,UTT,SEC)	0.19, 0.19, 0.19	↑
i_PER_UTT	0.19	↑
so_PER_(TOKEN,UTT,SEC)	0.18, 0.18, 0.16	↑
of_PER_SECOND	0.15	↑
a_PER_SECOND	0.15	↑
had_PER_UTT	0.13	↑
max $\Delta$ mfcc[6]	0.18	↑
Linear Regression slope F0	0.18	↓
max $\Delta$ mfcc[10]	0.16	↓

Table 3: Info gain valuesh

using more stilted language. Also, accommodation theory or entrainment may provide an explanation. Niederhoffer and Pennebaker (2002) discovered that conversation participants tend to mimic each others’ styles. Levitan et al. (2012) and Levitan et al. (2015a) showed that this behavior remains persistent through speech as well. For an extensive analysis of the representations and models, refer to our previous work (Katerenchuk et al., 2014).

## 5. Conclusion

We release a set of labels for the CALLHOME English telephone conversation corpus. The labels describe the relationships between the participants as friends or family members. This dataset should enable researchers to work on analyzing textual and acoustic information in conversations among friends or family. Understanding the patterns may enable researchers to use this knowledge and improve various NLP tasks. The labels are freely available for download at [https://github.com/dkaterenchuk/callhome\\_labels](https://github.com/dkaterenchuk/callhome_labels).

## 6. Acknowledgements

We would like to thank Min Ma, Michelle Morales, Rachel Rakov, Syed Reza, and Clayton Violand for their annotation work in this project.

This material is based on research sponsored by DARPA under agreement number FA8750-13- 2-0041. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

## 7. Bibliographical References

- Asami, T., Masumura, R., Yamaguchi, Y., Masataki, H., and Aono, Y. (2017). Domain adaptation of dnn acoustic models using knowledge distillation. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5185–5189. IEEE.
- Campbell, N. (2002). Recording techniques for capturing natural every-day speech. In *LREC*.
- Canavan, A., Graff, D., and Zipperlen, G. (1997). Call-home american english speech. Linguistic Data Consortium, Philadelphia.
- Carrascal, J., de Oliveira, R., and Cherubini, M. (2012). A Note Paper on Note-Taking: Understanding Annotations of Mobile Phone Calls.
- Chen, X., Tan, T., Liu, X., Lanchantin, P., Wan, M., Gales, M. J., and Woodland, P. C. (2015). Recurrent neural network language model adaptation for multi-genre broadcast speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.
- Chung, C. and Pennebaker, J. W. (2007). The psychological functions of function words. *Social communication*, pages 343–359.
- Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.
- Fourtassi, A., Schatz, T., Varadarajan, B., and Dupoux, E. (2014). Exploring the relative role of bottom-up and top-down information in phoneme learning. In *ACL (2)*, pages 1–6.
- Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth Annual Conference of the International Speech Communication Association*.
- Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., and Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44.
- Katerenchuk, D., Brizan, D. G., and Rosenberg, A. (2014). “Was that your mother on the phone?”: Classifying interpersonal relationships between dialog participants with lexical and acoustic properties. In *Fifteenth Annual Conference of the International Speech Communication Association*.
- Kendall, T. (2011). Corpora from a sociolinguistic perspective. *Revista Brasileira de Linguística Aplicada*, 11(2):361–389.
- Levitan, R., Gravano, A., Willson, L., Benus, S., Hirschberg, J., and Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 11–19. Association for Computational Linguistics.
- Levitan, R., Benus, S., Gravano, A., and Hirschberg, J. (2015a). Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*.
- Levitan, S. I., An, G., Wang, M., Mendels, G., Hirschberg, J., Levine, M., and Rosenberg, A. (2015b). Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 1–8. ACM.
- Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of japanese. In *LREC*.
- Mendels, G., Levitan, S. I., Lee, K.-Z., and Hirschberg, J. (2017). Hybrid acoustic-lexical deep learning approach for deception detection. *Proc. Interspeech 2017*, pages 1472–1476.
- Niederhoffer, K. G. and Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.
- Polzehl, T., Moller, S., and Metze, F. (2010). Automatically assessing personality from speech. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 134–140. IEEE.
- Ranganath, R., Jurafsky, D., and McFarland, D. (2009). It’s not you, it’s me: detecting flirting and its misperception in speed-dates. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 334–342. Association for Computational Linguistics.
- Rao, K. S., Kumar, T. P., Anusha, K., Leela, B., Bhavana, I., and Gowtham, S. (2012). Emotion recognition from speech. *International Journal of Computer Science and Information Technologies*, 3(2):3603–3607.
- Sanaullah, M. and Gopalan, K. (2013). Deception detection in speech using bark band and perceptually significant energy features. In *Circuits and Systems (MWSCAS), 2013 IEEE 56th International Midwest Symposium on*, pages 1212–1215. IEEE.
- Saon, G., Soltau, H., Nahamoo, D., and Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *ASRU*, pages 55–59.
- Schuller, B., Steidl, S., and Batliner, A. (2009). The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.
- Serizel, R. and Giuliani, D. (2017). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering*, 23(3):325–350.
- Siddique, F. B., Kampman, O., Yang, Y., Dey, A., and Fung, P. (2017). Zara returns: Improved personality induction and adaptation by an empathetic virtual agent. *Proceedings of ACL 2017, System Demonstrations*, pages 121–126.
- Stirman, S. W. and Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine*, 63(4):517–522.
- Yella, S. H., Anguera, X., and Luque, J. (2014). Inferring social relationships in a phone call from a single party’s speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4843–4847. IEEE.