Phonotactic Reconstruction of Encrypted VoIP Conversations: Hookt on fon-iks

Adam White, Austin Matthews, Kevin Snow, and Fabian Monrose

Presented By Corly Leung

Introduction

- Google Hangout, Skype, FaceTime
- Encrypting VoIP Packets
  - Variable-Bit-Rate for speech encoding
  - Length-preserving stream ciphers
    - Determine language spoken, identity, and presence of known phrases

Background

- Phonetic Models of Speech
  - Individual units of phones
    - Consonants vs Vowels
    - Characterize by articulatory processes
  - Alphabets for representing phones: International Phonetic Alphabet (IPA)
- Voice over IP
  - Audio encoded with an audio codec
    - Code Excited Linear Prediction
      - Excitation signal and Shape Signal

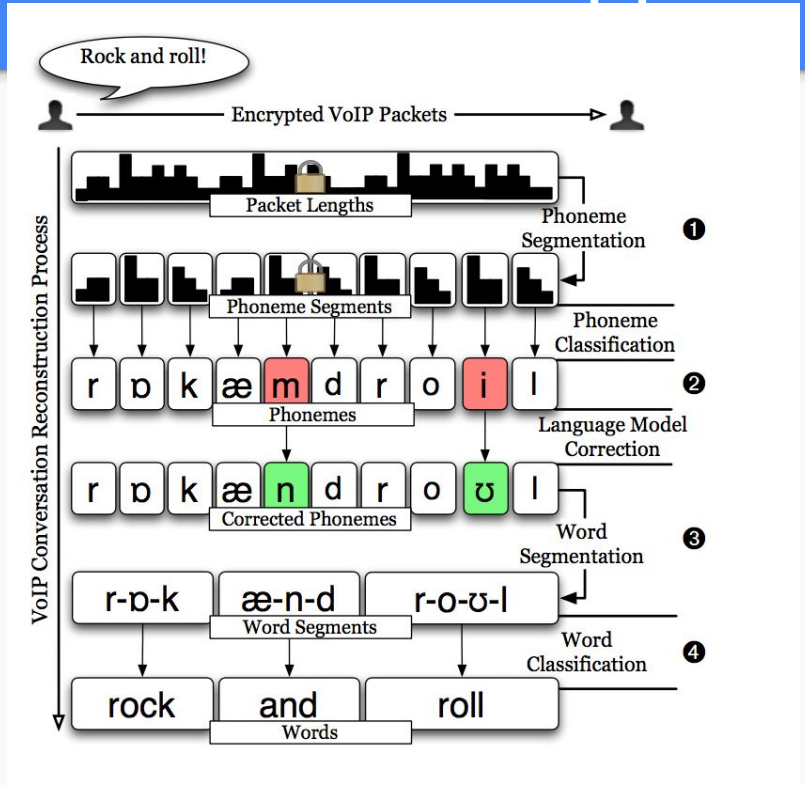
Related Works

- Traffic Analysis of Encrypted Network
- Encrypted VoIP calls to infer language and match to known phrases
- Silence suppression to identify speeches

Data and Adversarial Assumptions

- TIMIT Acoustic-Phonetic Continuous Speech Corpus
  - Collection of Speech with time-aligned word and phonetic transcripts
  - Encoded to Speex encoded
- Adversary
  - Sequence of Packet Lengths for an encrypted VoIP call
  - Knowledge of the language
  - Representative example of sequences for each phoneme
  - Phonetic dictionary

High Level Overview of Approach



Finding the Phoneme Boundaries

- Identify which packets represent a portion of speech containing boundary between phonemes.
- Maximum entropy modeling by maximizing p(w|v)
- Evaluation: Cross Validation with about 0.85 accuracy for n=1
  - n frames within boundary

Phoneme Segmentation Feature Templates	
1	size of frame w_i (i.e., the current frame size)
2	size of frame w_{i-1} (i.e., the previous frame size)
3	size of frame w_{i+1} (i.e., the next frame size)
4	bigram of sizes for frames w_{i-1}, w_i
5	bigram of sizes for frames w_i, w_{i+1}
6	trigram of sizes for frames w_{i-1}, w_i, w_{i+1}
7	sequence of frame sizes since the last hypothesized boundary
8	number of frames since since the last hypothesized boundary

Classifying the Phonemes

- Classification problem of various phonemes
- Context dependent
  - Maximum entropy modeling: model only parameters of interest
- Context independent
  - Profile hidden Markov modeling: model entire distribution over examples
- Bayesian inference to update posterior given by maximum entropy classifier with evidence by HMM
- Enhancing Classification using Language Modeling
- Evaluation: 77% context dependent, 67% context independent vs 69% human

Segmenting Phoneme Streams Into Words

- Identify likely word boundaries
  - Insert potential word breaks into sequence of phonemes
  - Pronunciation dictionary to find valid word matches
    - [ɪn ɔɪli ɹæg] (‘an oily rag’)
    - [ɪn ɔɪl i ɹæg] (‘an oil E. rag’)
    - [ɪn ɔ ɪl i ɹæg] (‘an awe ill E. rag’)
- Evaluation: Precision 73% and Recall 85%

Identifying Words via Phonetic Edit Distance

- Convert Subsequences of Phonemes into English Words
  - Phonetically based alignment method
    - Distance between two vowels/ consonants by rounding, backness, height or voice, manner, and place of articulation
  - Phonetic distance between sequence and each pronunciation in dictionary
    - Homophones (eight vs ate)
      - Word and part of speech model

Overall Evaluation

- Speaker independent model
- Content-dependent
  - Multiple utterance of particular sentence
  - Scoring around 0.67 and 0.9 with 0.5 being understandable
- Content-independent
  - All TIMIT utterances
  - 0.45 average

Measuring Confidence

- Close pronunciation matches are more likely to be correct than distant matches
- Mean of probability estimates of each word in hypothesized transcript
- Forgoing less confidence words

Mitigations

- Varying frame based per packet
- Packets are observed in correct order
- Relatively large block sizes
- Constant bit-rate codecs
- Drop or packets

Discussion

- What are the key contributions of the paper?
- How practical is the attack?
- Are the mitigations sufficient?