

# Uncovering Spoken Phrases in Encrypted VOIP Conversations

Trent Kalisch-Smith

Charles V. Wright  
Lucas Ballard  
Scott E. Coull  
Fabian Monroe  
Gerald M. Masson

# Outline



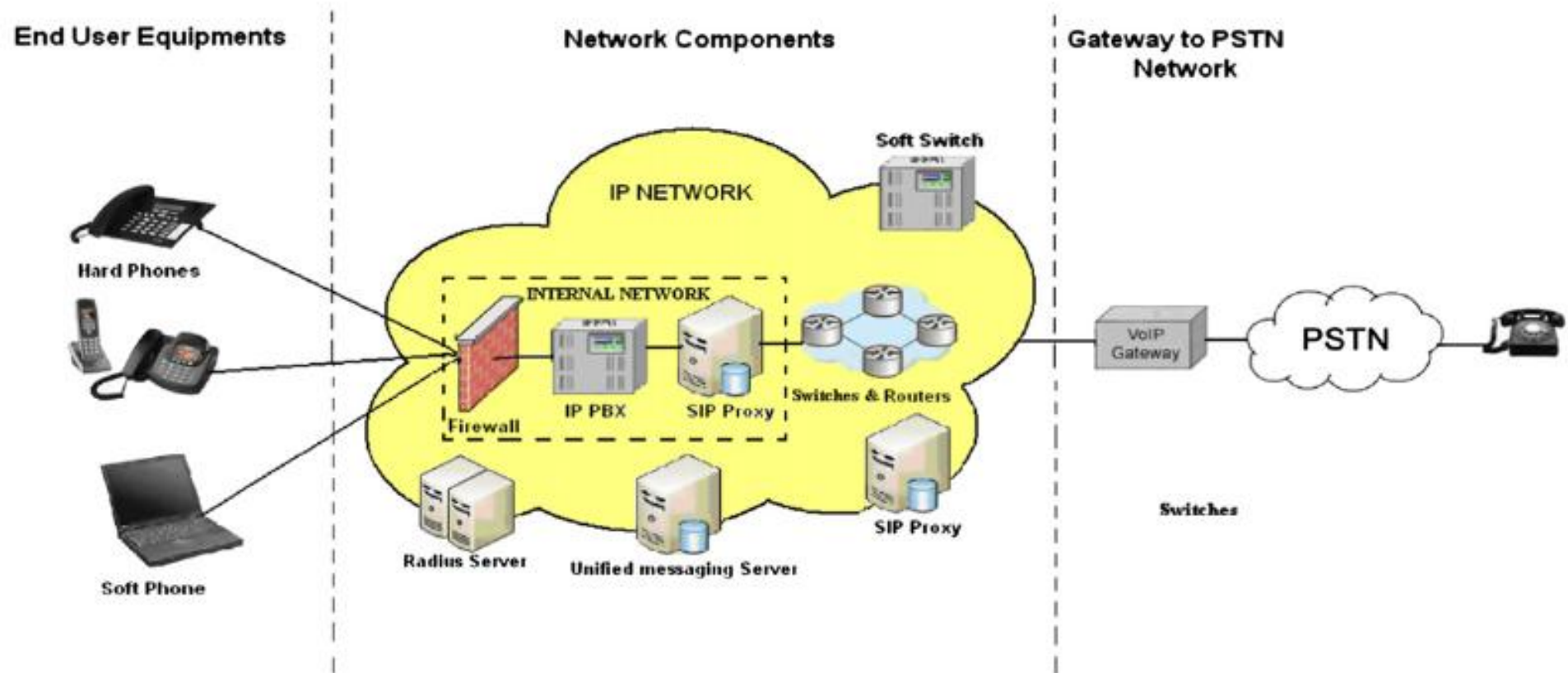
- Overview of how VOIP works
- Speech Processing in VOIP
- The problem of variable bit rate encoding
- How an attacker can exploit this vulnerability to uncover words and phrases
- Example of an attack
- Performance
- Mitigation
- Threat posed to current systems

# VOIP

- Voice over Internet Protocol
- Communicating via the Internet rather than PSTN
- Separates connection set up and transmission of data.
  - Control Channel e.g. SIP, H.323
  - Data transmitted using Real-time Transfer Protocol
  - Encrypted with SRTP which uses a length preserving stream cipher

# Overview of VOIP

- Voice over IP



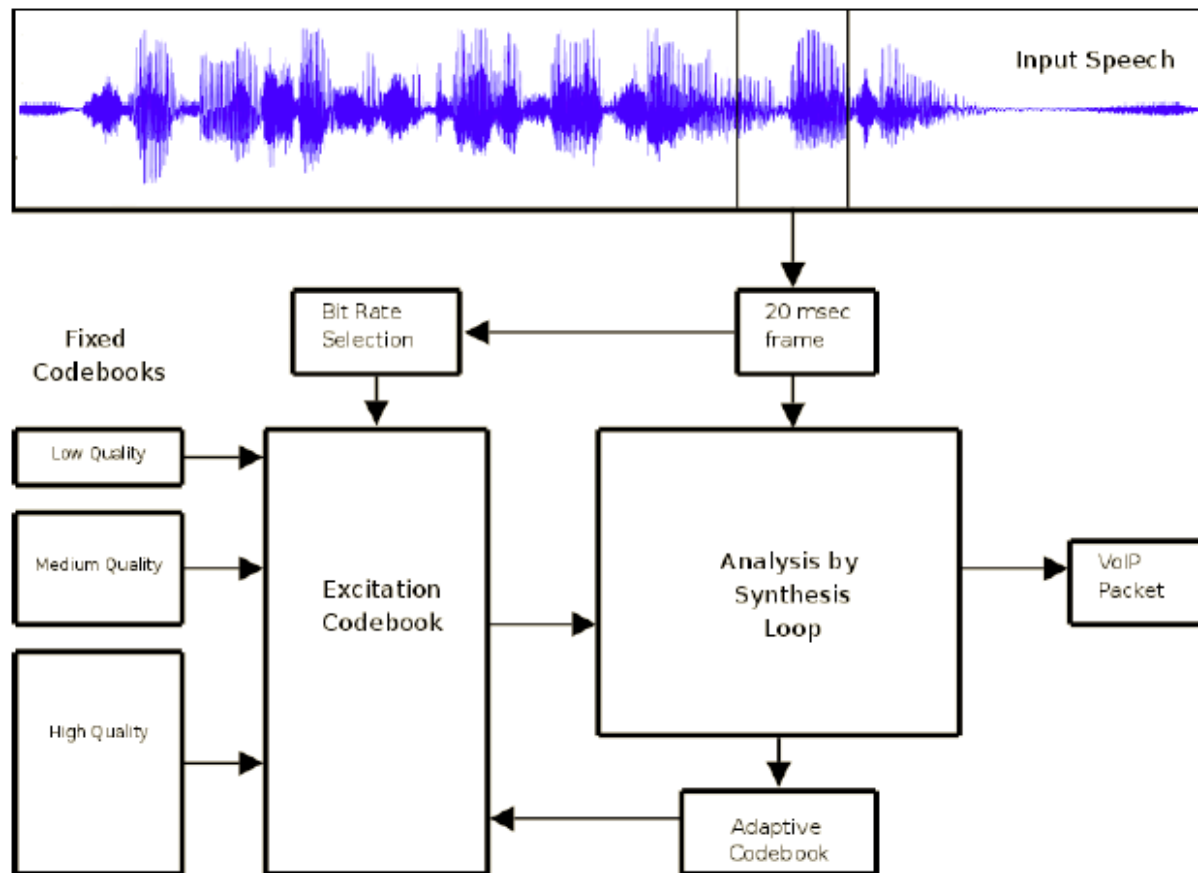
# Turning Voice into Packets

- User's voice is typically sampled at 8000 or 16000Hz.
- Audio codec takes stream of audio samples and compresses it into a digital form.
  - Code Excited Linear Prediction
    - Speex (variable bit rate mode enabled)

# Code Excited Linear Prediction (CELP)

- Keeps a codebook of audio vectors
- Reads the original sampled audio and performs a brute force search over the codebook to output the vector that most closely resembles the original sample
- The payload of the packet consists of
  - The index of the best fitting entry in the codebook
  - Linear predictive coefficients
  - The gain
- The encoder may optionally adaptively choose the bit rate for each packet to balance quality and bandwidth

# Speex's CELP encoder



# The Problem

- Compressing packets with a variable bit rate before encrypting them with a length preserving stream cipher leaks information.
- The choice of bit rate is based on packet's payload

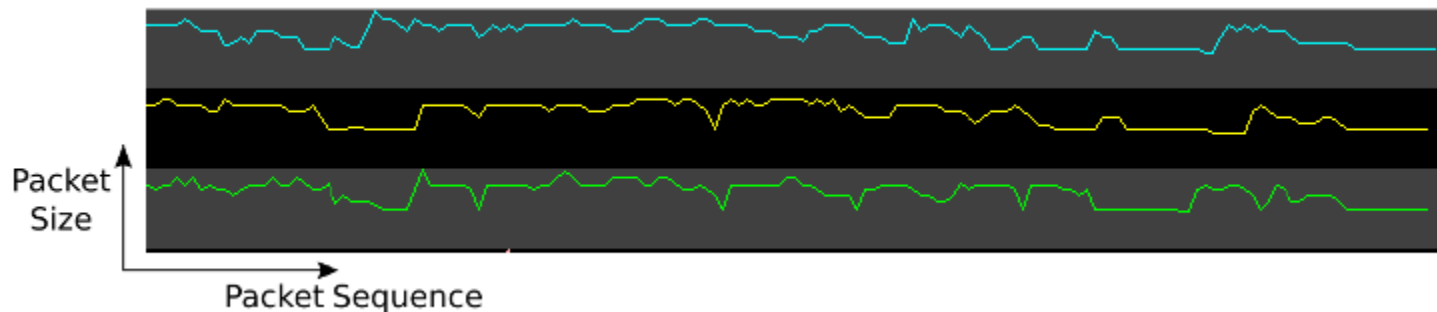
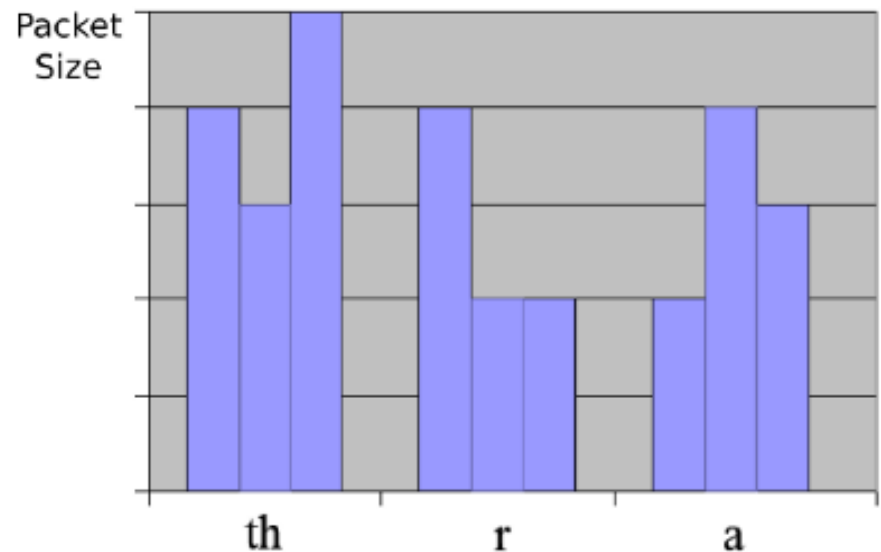


Figure2: The same sentence spoken three times and compressed using variable bit rate encoding.

# Phonemes

- Smallest segments of a sound that together form the words in our language
- 40-60 in total
- Each are made up of a different combination of packet sizes when encoded



How can we find words or phrases in a stream of packets by only looking at their size?

# Profile Hidden Markov Model

- Used in speech recognition and biometric matching
- Works well for finding patterns in data with high variability
- Does not require knowledge of the speaker or any examples of the audio produced by them speaking the target words or phrases.

# The Strategy

- Build a Profile HMM to recognise phonemes within the packet stream using a wide variety of pronunciations
- Search a packet stream and calculate possible sequences of phonemes using the HMM
- Apply Viterbi decoding on the stream of phonemes to find the most likely sequence of phonemes that match the stream of packets.

# Assumptions for the Attack

- The VOIP call is made using a VBR audio codec and a length preserving stream cipher
- The language used by the callers is known prior to eavesdropping
- The attacker has statistics defining phonemes and their corresponding packet lengths

### 1. Gather Training Data

Pronunciation  
Dictionary



Phoneme Examples



### 2. Encode Phonemes into VoIP Packets

a' →

r →

th →

⋮  
⋮  
⋮

### 3. Synthesize Phrase Training Data

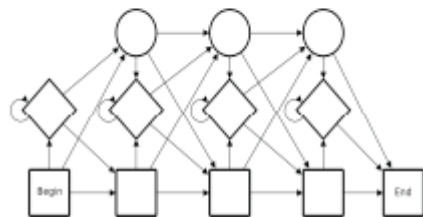
"The target ..."



th e t a' r g ...



### 4. Train HMM with Synthetic Data



### 5. Spot the Phrase in VoIP Packets



"The target ..."

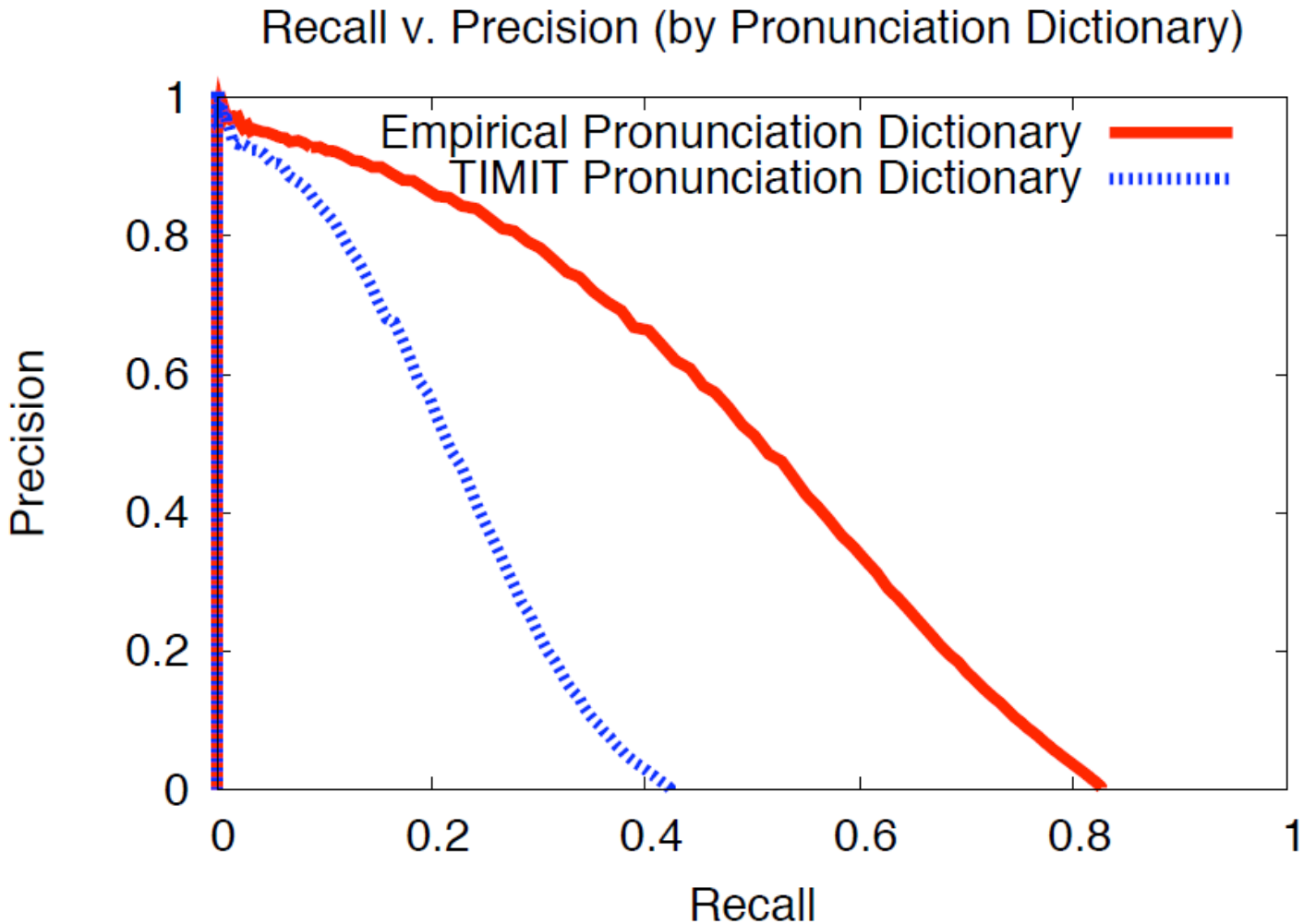
# Evaluating the Effectiveness of the Attack

- Collected 122 target sentences to be searched for from the TIMIT corpus
  - 6300 sentences
  - 630 speakers
  - Wide range of English accents
  - Time aligned phonetic transcriptions
- 1<sup>st</sup> attempt used the pronunciation dictionary included with TIMIT.
- 2<sup>nd</sup> attempt used a combination of the TIMIT dictionary, speech examples and two other speech corpuses to form an Empirical Pronunciation Dictionary.

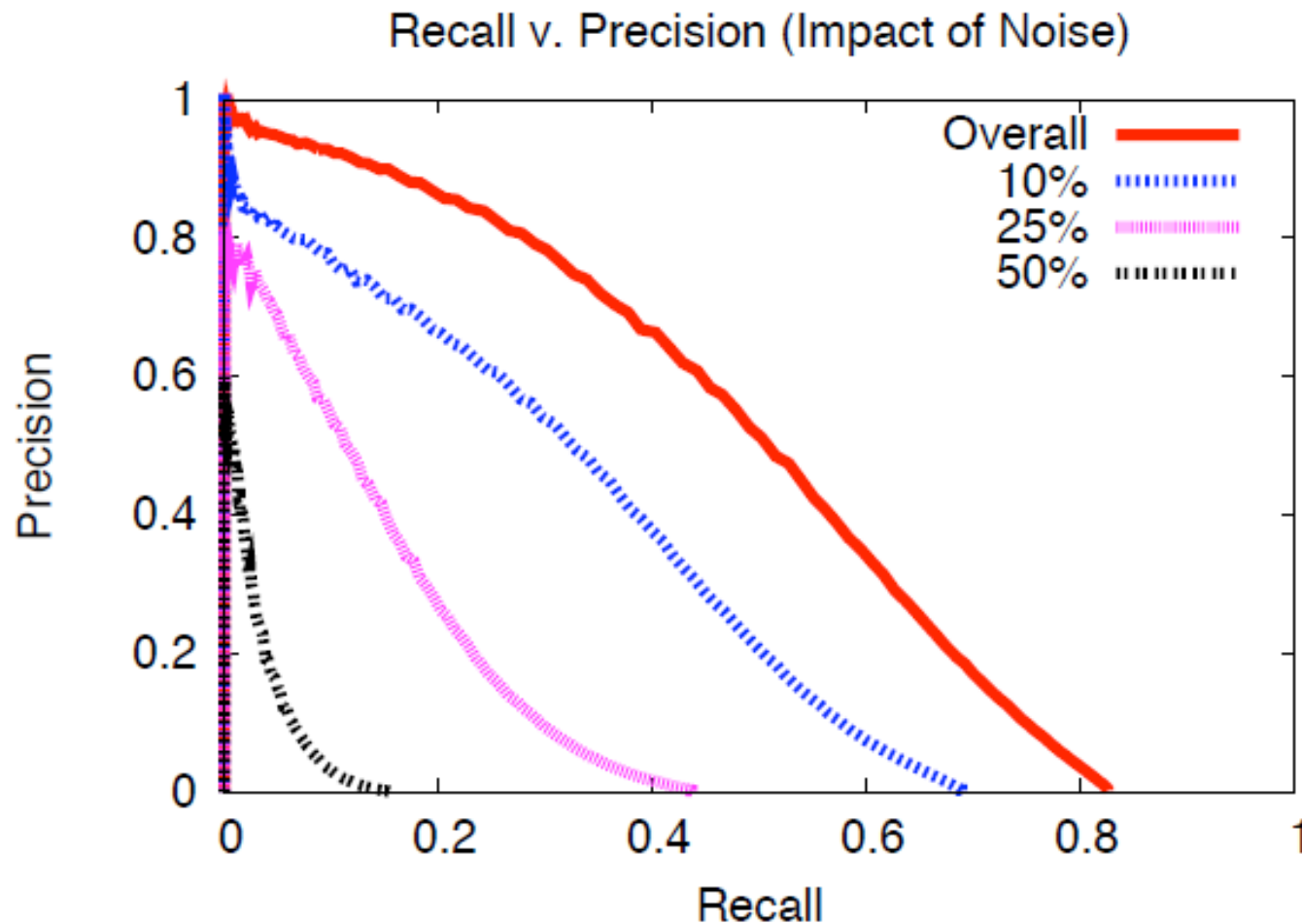
# Performance Metrics

- *Precision* = probability that a reported match was completely correct.
- *Recall* = the probability that the algorithm found the phrase if the phrase was contained within the ciphertext

# Performance



# Robustness to Noise

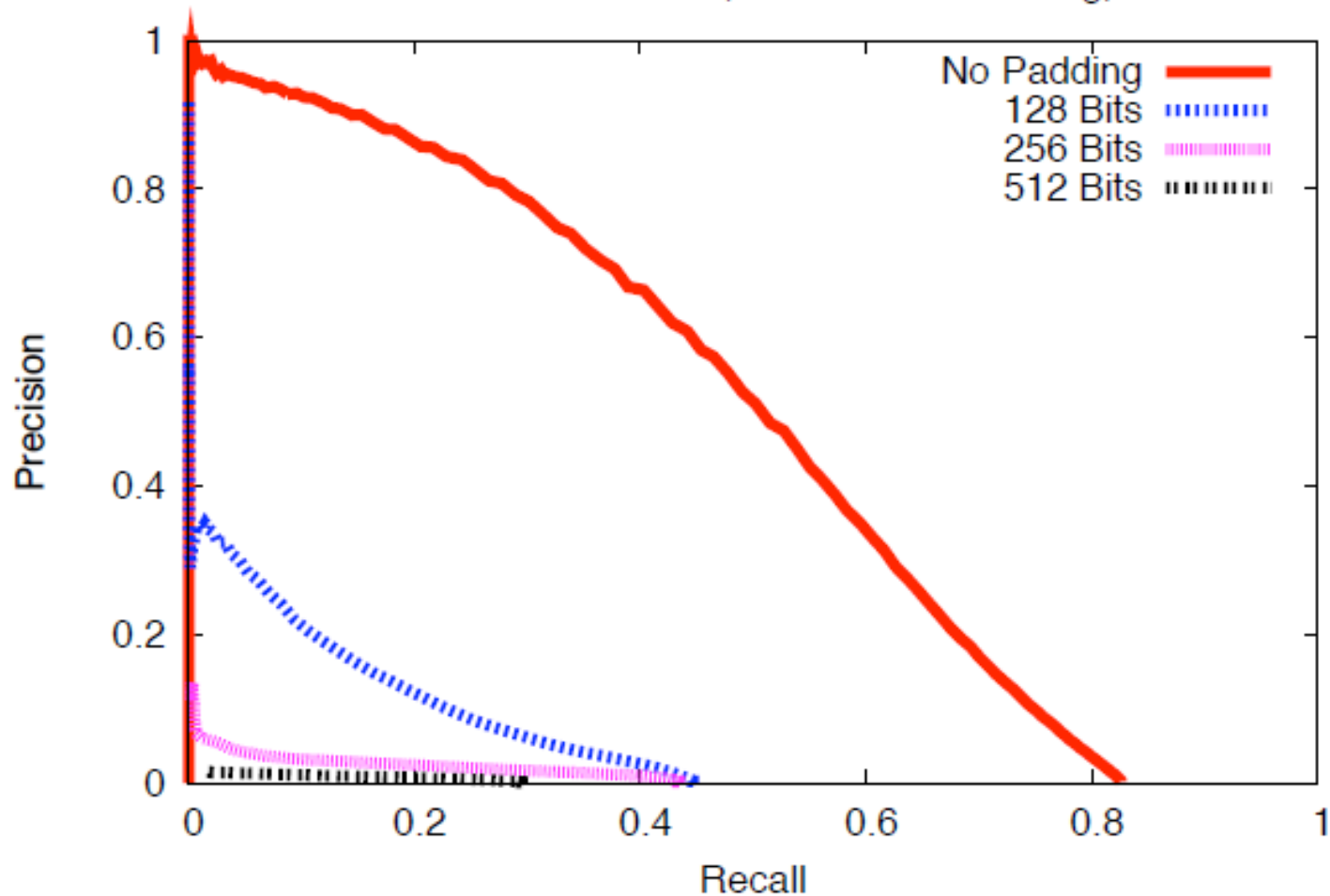


# Mitigation

- Pad packets to a common length or add random padding before encryption
- 128 bit blocks add an extra 8.8% overhead
- 256 bit blocks add an extra 16.5% overhead
- 512 bit blocks add an extra 30.8% overhead

# Before and After Padding

Recall v. Precision (The Effect of Padding)

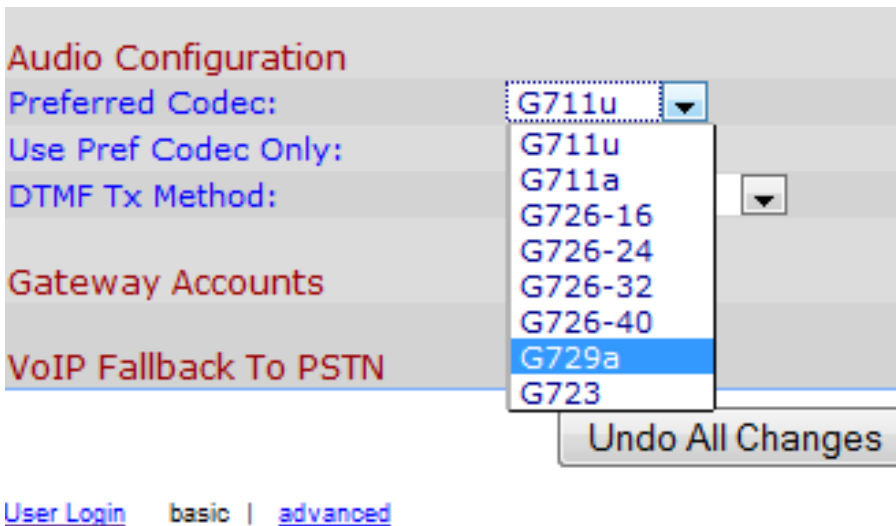


What else can be done to mitigate against this attack?

# Is it safe to Skype?

- Skype uses several closed source audio compressors
- G729a
  - Fixed bitrate
  - All packets same frame size
  - Secure against this type of attack
- SILK
  - Developed by Skype and released in 2009
  - Variable bit rate!
  - Skype uses block cipher (AES) to encrypt so blocks will be constant length.

# Check your VOIP



The screenshot shows a web-based configuration interface for VoIP. The 'Preferred Codec' dropdown menu is open, displaying a list of codecs: G711u, G711a, G726-16, G726-24, G726-32, G726-40, G729a (highlighted), and G723. Other visible settings include 'Use Pref Codec Only:', 'DTMF Tx Method:', 'Gateway Accounts', and 'VoIP Fallback To PSTN'. At the bottom, there are links for 'User Login', 'basic', and 'advanced', and a copyright notice for Cisco.

The following audio codecs use a type of VBR encoding

- QCELP
- Speex
- G729.1
- GSM also uses CELP

# Risks of being exposed to this attack

- There are many alternatives to CELP
  - Modified discrete cosine transform
- Not many CELP varieties actually use variable bit rate
  - VBR mode must be enabled in Speex
- Not all technologies use length preserving stream ciphers

# Conclusion

- VOIP techniques that combine a variable bit rate audio encoder with a length preserving stream cipher leak information
- Study showed that it is possible to find about approximately 50% of commonly spoken phrases.
- This is an easy trap to fall into but can be guarded against simply

# Questions

# References

- Spittka, et al, 2010 *RTP Payload Format for SILK*, Available from: <http://tools.ietf.org/html/draft-spittka-silk-payload-format-00#section-8>
- Wright, C. V., Ballard, L., Coull, S. E., Monroe, F., and Masson, G. M. 2008. *Spot Me if You Can: Uncovering Spoken Phrases in Encrypted VoIP Conversations*. Proceedings of the 2008 IEEE Symposium on Security and Privacy (May 18 - 21, 2008). SP. IEEE Computer Society, Washington, DC, 35-49.