Silence Detection in Secure P2P VoIP Multiconferencing

JOSE-VICENTE AGUIRRE[1], RAFAEL ÁLVAREZ[2], JULIA SÁNCHEZ[3],
ANTONIO ZAMORA[4]
Dpt. of Computer Science and Artificial Intelligence
University of Alicante
Campus de San Vicente, Ap. Correos 99, 03080 Alicante
SPAIN
jaguirre@dccia.ua.es[1], ralvarez@dccia.ua.es[2], jsanchez@dccia.ua.es[3], zamora@dccia.ua.es[4]

This work was partially supported by the Spanish grants GV06/018 and MTM2005-05759

Abstract: - We analyze the impact of silence detection applied to secure P2P VoIP multiconferencing. Under the assumption that the number of simultaneous conversations is small regardless of the amount of participants we propose considering simple audio broadcasting as the basis of the multiconference scheme and study different alternatives regarding multiplexing, cryptographic primitives and implementation issues.

Key-Words: - Multiconferencing, Broadcasting, VoIP, Audio, Security, Silence detection, P2P, Multiplex.

1 Introduction

Most P2P audio multiconferencing systems focus on the worst case in order to achieve adequate performance and fair load sharing for all cases.

The characteristics of human conversation make silence detection an effective way of eliminating background noise and reducing bandwidth requirements, since most of the time there will only be one active speaker even if the number of participants is very high.

In this scenario, considering that audio multiconferencing will degenerate into a simple broadcast a significant amount of the time, is a suitable approach, extending this basic broadcasting to manage audio stream multiplexing.

2 Preliminaries

We present here several key concepts which are necessary for the adequate understanding of the techniques introduced in the rest of the paper.

2.1 Mixing

Mixing audio signals is the process of combining them into a new audio signal that sounds like all of them were played together. In the digital domain it is realized by adding each sample together, so for 2 signals

$$M_i = A_i + B_i,$$

where *i* corresponds to the sample index.

Conversely for *n* signals

$$M_i = S_i^1 + S_i^2 + \cdots + S_i^n.$$

The main problem with digital mixing is that of the limited headroom: if we add too many loud signals we run the risk of exceeding the maximum values that can be represented with a certain bit depth, resulting in audio distortion. This problem can usually be minimized employing floating point arithmetic.

2.2 Gating

A noise gate is a device or algorithm that shuts down an audio signal when the level is below a certain threshold. In this way, when a microphone is not being used, the noise gate can reduce the volume and virtually turn the microphone off, preventing any noise or undesired weak signal from being added to the mix.

Noise gates have several parameters:

- *Threshold*, or the level at which the gate begins to close or open.
- *Attack*, or the time it takes for the gate to open.
- *Release*, or the time it takes for the gate to close.

The attack and release values can be constant or program dependent, varying in time as a function of the audio being processed.

2.3 Cryptography

Secure VoIP multiconferencing is based on cryptographic primitives, generally employing asymmetric cryptosystems (RSA [4], Diffie Hellman [1] among others) for session key exchange and symmetric ciphers (AES, RC4, DES, etc., see [2], [5] and [6]) for session data encryption.

In the case of symmetric cryptosystems, there are two classes, block ciphers and stream ciphers, with different properties and applications.

2.3.1 Block ciphers

A block cipher is a symmetric key cipher which operates on groups of bits of fixed length, termed blocks, with an unvarying transformation.

In encryption, they take a block of plaintext and produce a block of the same size of ciphertext. In decryption, they produce the corresponding block of plaintext taking the ciphertext block as input. The exact transformation is determined by the secret key that is taken as input to certain parts of the algorithm.

2.3.2 Stream ciphers

A stream cipher is a symmetric cipher in which the plaintext elements are encrypted one at a time, and in which the transformation of successive elements varies as the encryption or decryption process progresses. Plaintext elements are generally single bits or bytes.

Unlike block ciphers, which are generally based on the Feistel scheme, stream ciphers are mostly based on the Vernam cipher.

In encryption, stream ciphers perform a *XOR* operation between a pseudorandom sequence (the key stream) and the plaintext obtaining the ciphertext. In decryption, the *XOR* operation is performed between the ciphertext and the same random sequence in order to obtain the original plaintext.

The key is generally the first state (the seed of the pseudorandom sequence generator).

3 Description

Designing P2P VoIP multiconferencing systems implies, usually, considering the worst case in order to guarantee adequate performance in all cases. This generally means that we have to consider that all participants are speaking at the same so all the audio is routed and mixed correctly.

We propose, as an alternative approach, to consider the ideal case and then study ways to manage the other, not so optimal, cases.

3.1 Silence Detection

Silence detection, sometimes called activity or on/off detection, has been extensively applied to the subject of VoIP communication (see [3] and [7]). It is performed by a noise gate device to

It has two main benefits: first, by not transmitting audio data that has no significant content, we can greatly reduce bandwidth and computational requirements; second, adding multiple background noises can be a very significant problem in audio multiconferencing, since the number of silent participants is potentially very high.

There are generally two ways of performing silence detection:

- *Frame based*, so that if a frame has any content (even if it is very little) it is transmitted and if it has absolutely no content it is discarded.
- *Stream based*, so that only the content portions of the audio are transmitted, effectively eliminating any non content audio. This requires variable sized frames.

3.2 The ideal case

In P2P multiconferencing the best case (besides the trivial case of every participant being silent) is to consider that there is only one participant talking.

This, effectively, constitutes a simple audio broadcasting in which we have two types of nodes:

- *Emitters*, nodes that are speaking or generating audio content.
- *Relays*, nodes that are silent so they just playback audio and transmit it to other nodes.

As shown in Fig. 1, emitters (marked as E nodes) generate audio and relays (marked as R nodes) transmit it to others nodes. In order to keep synchronization, the emitter waits until an acknowledgement (marked as ready) is received to proceed with broadcasting.

If we work under the reasonable assumption that, during a multiconference, most of the time there will be only one active participant simultaneously, then the idea of using broadcasting as the main case for P2P multiconferencing is a suitable approach. This assumption stems from the idea that human beings are capable of understanding a limited number of simultaneous conversations, regardless of the number of participants.

In the following, we study some alternatives to manage additional audio streams over the basic broadcasting concept.

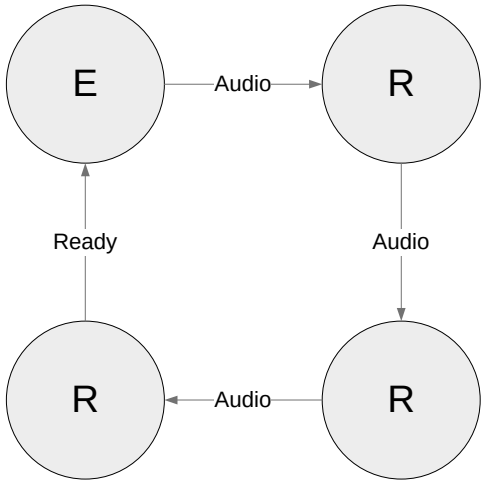

Fig. 1. Basic audio broadcasting scheme.

3.3 Broadcast multiplexing

In order to provide multiconferencing over the broadcasting scheme shown in Fig. 1, the trivial approach would be to, simply, perform several broadcasts at the same time. This case is shown in Fig. 2.

The result is that the bandwidth requirements are effectively multiplied by the number of emitters. These would result in dramatic increases in bandwidth utilization even for a small number of simultaneous emitters.

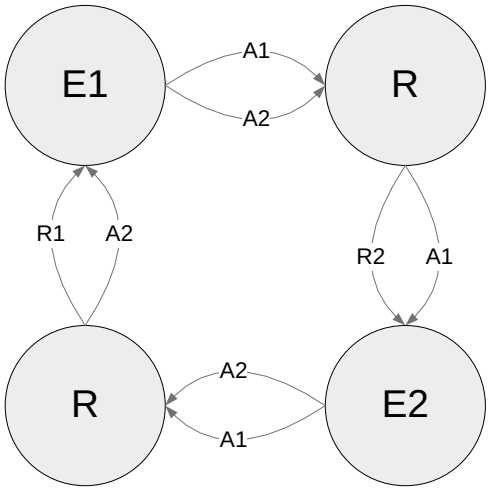

Fig. 2. Broadcast multiplexing without mixing.

A more reasonable approach would be to consider submixing audio streams in order to reduce the number of required packet transmissions. This approach is shown in Fig. 3 for 2 emitters and in Fig. 4 for 3 emitters.

In this case, the audio is mixed sequentially and then distributed to all nodes. This provides a significant performance improvement over simple broadcast multiplexing and at the same time is conveniently scalable.

We can observe in Table 1 that the total number of packet transmissions for the broadcast multiplexing with mixing approach does not increase significantly even for high multiplex factors which are, nevertheless, not expected under normal conditions. In Table 1, *n* is the total number of nodes and *m* is the multiplex factor (number of emitters).

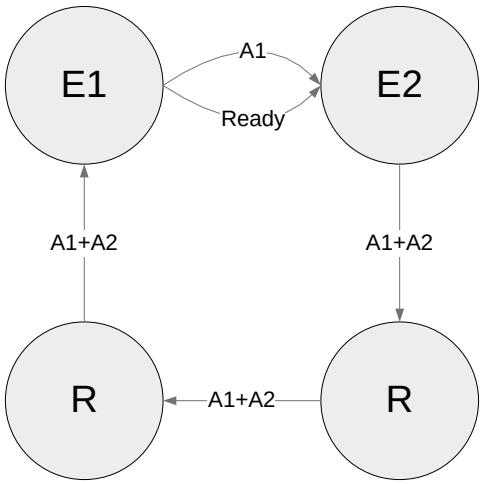

Fig. 3. Broadcast multiplexing with mixing for *m*=2.

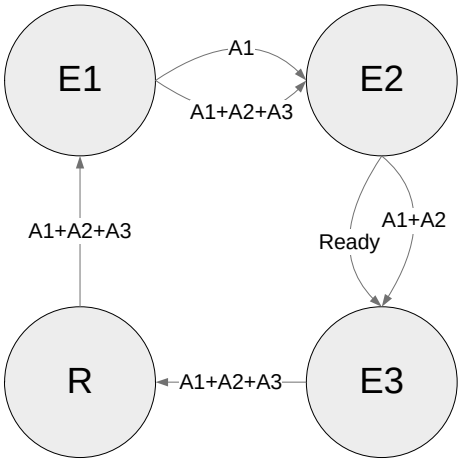

Fig. 4. Broadcast multiplexing with mixing for *m*=3.

Scheme	Packets
Simple broadcast	$n-1$
Multiplex, no mixing	$n \times m$
Multiplex, mixing	$n+m-2$

Table 1. Audio packet transmissions for different schemes.

4 Analysis

Although stream based silence detection is better at reducing the amount of non content audio transmitted, in a secure VoIP multiconferencing system, the choice of symmetric cipher plays an important role because using variable sized packets is not optimal while employing a block cipher since padding would be necessary in order to reduce the total audio delay. A stream cipher is more convenient in this application.

Key management is also a topic that requires certain analysis. Using only a key per multiconference session is easier but employing different keys per link would be more secure and would permit advanced features like subgroups at the cost of additional computation and protocol complexity.

Another consideration is that the nodes must be reorganized so that emitters appear first in the transmission chain; this requires additional control information to be sent in order to manage node connections. This restriction can be relaxed but performance would be suboptimal.

5 Conclusion

We have analyzed the impact of silence detection in secure VoIP multiconferencing, proposing an alternative approach focused on the idea that the number of simultaneous conversations is relatively small regardless of the amount of participants.

This allows considering multiconferencing as a simple broadcasting scheme that can be extended to perform audio stream multiplexing.

Applying silence detection to secure P2P VoIP multiconferencing is an effective way to reduce bandwidth and computational requirements and, therefore, increase the total number of possible participants.

The proposed scheme can be extended to non sequential node connections like trees or graphs.

References:

[1] Diffie, W., Hellman, M., New directions In Cryptography, IEEE Trans. Information Theory, Vol. 22, 1976, pp. 644-654

[2] Menezes, A., van Oorschot, P., Vanstone, S., Handbook of Applied Cryptography, CRC Press, Florida, 2001

[3] Prasad, R. V., Sangwan, A., Jamadagni, H. S., Chiranth, M. C., Sah, R., Gaurav, V., Comparison of Voice Activity Detection Algorithms for VoIP, Proc. Int. Symp. Computers and Communications (ISCC), 2002, pp. 530-535

[4] Rivest, R., Shamir, A., Adleman, L., A Method for Obtaining Digital Signatures and Public Key Cryptosystems, ACM Communications, Vol. 21, 1978, pp. 120-126

[5] Schneier, B., Applied Cryptography Second Edition: protocols, algorithms and source code in C, John Wiley and Sons, New York, 1996

[6] Stallings, W., Cryptography and Network Security: Principles and Practice. Third Edition, Prentice Hall, New Jersey, 2003

[7] Wenyu, J., Schulzrinne, H., Analysis of On-Off Patterns in VoIP and their Effect on Voice Traffic Aggregation, Proc. IEEE Int. Conf. Computer Communication Networks, 2000, pp. 82-87