TCVN TIÊU CHUẨN QUỐC GIA

TCVN 9804:2013

Xuất bản lần 1

CHẤT LƯỢNG DỊCH VỤ VIDEO THOẠI TỐC ĐỘ THẤP SỬ DỤNG CHO TRAO ĐỔI NGÔN NGỮ KÝ HIỆU VÀ ĐỌC MÔI THỜI GIAN THỰC

Sign language and lip-reading real-time conversation using low bit-rate video communication

HÀ NỘI - 2013

Mục lục

Lời nói đầu

TCVN 9804 : 2013 được xây dựng trên cơ sở tham khảo Khuyến nghị họ H – Phần phụ 1 của Liên minh Viễn thông quốc tế ITU-T.

TCVN 9804 : 2013 do Viện Khoa học Kỹ thuật Bưu điện xây dựng, Bộ Thông tin và Truyền thông đề nghị, Tổng cục Tiêu chuẩn Đo lường Chất lượng thẩm định, Bộ Khoa học và Công nghệ công bố.

Chất lượng dịch vụ video thoại tốc độ thấp sử dụng cho trao đổi ngôn ngữ ký hiệu và đọc môi thời gian thực

Sign language and lip-reading real time conversation using low bit-rate video communication

1 Phạm vi áp dụng

Tiêu chuẩn này áp dụng cho trao đổi ngôn ngữ ký hiệu và đọc môi, bao gồm các đặc tính cần thiết của một hệ thống truyền thông video cho hội thoại giữa người và người sử dụng ngôn ngữ ký hiệu và đọc môi có hoặc không có thoại âm thanh.

Tiêu chuẩn này quy định yêu cầu hiệu năng cần được đáp ứng để đảm bảo cho cuộc hội thoại thành công.

Tiêu chuẩn này mô tả cách đánh giá hiệu năng chất lượng dịch vụ video thoại tốc độ thấp sử dụng ngôn ngữ ký hiệu và đọc môi.

2 Tài liệu viện dẫn

Tài liệu viện dẫn sau rất cần thiết cho việc áp dụng tiêu chuẩn này. Đối với các tài liệu viện dẫn ghi năm công bố thì áp dụng phiên bản được nêu. Đối với các tài liệu viện dẫn không ghi năm công bố thì áp dụng phiên bản mới nhất, bao gồm cả sửa đổi, bổ sung (nếu có).

ITU-T Recommendation G.114 (1996), One-way transmission time. *(Thời gian truyền dẫn một chiều);*

ITU-T P.931 (12/98) Multimedia communications delay, synchronization and frame rate measurement *(Đo tốc độ khung, đồng bộ và trễ truyền thông đa phương tiện);*

3 Thuật ngữ và định nghĩa

3.1

Khung (frame)

Một hình ảnh hoàn chỉnh trong sản xuất video được gọi là một "khung". Trong một số hệ thống, các khung hình được xây dựng bởi hai nửa hình ảnh, mỗi nửa có chứa các thông tin trong khung. Những nửa hình ảnh đó được gọi là "vùng".

3.2

Độ phân giải (resolution)

Độ phân giải là độ sắc nét của hình ảnh thể hiện qua số dòng và số cột của màn ảnh hay số phần tử hình ảnh trên một đơn vị diện tích.

3.3

Độ tương phản (contrast)

Độ tương phản là sự khác nhau về màu sắc giữa hình ảnh và nền ảnh.

3.4

Ngôn ngữ ký hiệu (sign language)

Ngôn ngữ ký hiệu được biểu hiện qua các cử động, vị trí của tay, mắt, miệng, mặt và cơ thể.

3.5

Đọc môi (lip-reading)

Ngôn ngữ đọc môi được thể hiện qua cử động của khuôn mặt. Thông thường đọc môi được hỗ trợ bởi tiếng nói. Trong các trường hợp khác nó được sử dụng cùng với ngôn ngữ ký hiệu. Có một số người khiếm thính không sử dụng ký hiệu mà chỉ sử dụng ngôn ngữ đọc môi.

3.6

Khung lặp (repeated frame)

Khung lặp là một khung hình đầu ra không được phân biệt với các khung trước nó trong chuỗi (khi các khung chuỗi đầu vào tương ứng có sự khác biệt rõ ràng).

3.7

Khung tích cực (active frame)

Khung tích cực (khung không lặp) là một khung hình đầu ra được phân biệt với các khung trước nó trong chuỗi (khi các khung chuỗi đầu vào tương ứng có sự khác biệt rõ ràng).

4 Chữ viết tắt

CIF	Kích cỡ khổ trung gian gồm 288 dòng x 352 điểm	Common Intermediate Format
QCIF	Kích cỡ một phần khổ trung gian gồm 144 dòng x 176 điểm	Quarter Common Intermediate Format
SQCIF	Kích cỡ một phần khổ trung gian gồm 96 dòng x 128 điểm	Sub Quarter Common Intermediate Format
fps	Số khung trên một giây; số ảnh ảnh trên một giây.	Frame per second

5 Các yêu cầu cơ bản của máy điện thoại thấy hình sử dụng để trao đổi ngôn ngữ ký hiệu và đọc môi

5.1 Các yêu cầu độ phân giải thời gian

Cả ngôn ngữ ký hiệu và đọc môi đều yêu cầu mô phỏng trực quan tốt các cử động. Một hệ thống mô phỏng chuyển động với các bức tranh phân bố đều, cần tuân theo những đặc tính sau đây:

- Tốc độ 20 khung hình trên giây (fps) phù hợp với ngôn ngữ ký hiệu và đọc môi;

- Với một số trường hợp, có thể sử dụng tốc độ khung từ 12 fps và cao hơn;

- Đối với đọc môi, khi sử dụng nhận thấy độ dốc tăng khi tốc độ khung tăng tới 15 fps. Lớn hơn 15 fps nếu vẫn cứ tiếp tục tăng thì hình ảnh hiển thị sẽ khó đọc được;

- Khả năng sử dụng rất bị hạn chế khi tốc độ khung hình nằm trong khoảng 8 - 12 fps, với suy giảm lớn về khả năng tiếp nhận hoặc tốc độ;

- Tốc độ khung dưới 8 fps không được sử dụng cho ngôn ngữ ký hiệu hoặc đọc môi.

5.1.1 Đánh vần bằng tay

Các yêu cầu về độ phân giải thời gian của ngôn ngữ ký hiệu được hình thành trong trường hợp đánh vần bằng tay. Đánh vần bằng tay là một kỹ thuật trong đó mỗi chữ cái tương ứng với một cử chỉ bằng tay duy nhất. Cách đánh vần bằng tay ở các nước khác nhau là khác nhau. Đánh vần được thực hiện bằng cách biểu diễn các cử chỉ bằng chuỗi các hành động (bằng tay) nhanh để hình thành các từ. Các từ được đánh vần thường là tên hoặc các danh từ riêng khác mà các ký hiệu chính của ngôn ngữ ký hiệu không có. Đánh vần bằng tay rất nhanh và thường sử dụng 10 chữ cái (hoặc dấu) trên giây. Với những chữ cái cần thể hiện chính xác, cần ít nhất 2 hình để biểu diễn một chữ cái. Với các từ khác, đánh vần bằng tay rõ nét yêu cầu ít nhất 20 khung trên giây.

5.1.2 Ký hiệu chung

Đánh vần bằng tay chỉ là một phần của ngôn ngữ ký hiệu. Phần lớn ngôn ngữ ký hiệu được thực hiện bằng các ký hiệu cho các khái niệm hoàn chỉnh, các câu không hoàn chỉnh, ngữ pháp và các danh từ thông thường. Có rất nhiều ngôn ngữ ký hiệu trên thế giới. Trong quá trình sử dụng ký hiệu nói chung, các cử động tay nhanh kết hợp với những cái chớp mắt ngắn mang thông tin về ngữ pháp. Trong nhiều trường hợp, các yêu cầu độ phân giải thời gian tương tự với những yêu cầu cho đánh vần bằng tay.

5.1.3 Đọc môi

Yêu cầu cho đọc môi có thể được tính từ tốc độ âm vị của thoại thông thường. Tốc độ thông thường là 10 âm vị trên giây. Yêu cầu tốc độ tối thiểu là 20 hình trên giây để cho phép người xem đọc được âm vị rõ ràng.

5.1.4 Khả năng thích ứng

Trong cả hai trường hợp đọc môi và ngôn ngữ ký hiệu, tốc độ tạo ra ngôn ngữ có thể được giảm bớt theo ý muốn. Điều đó giải thích tại sao có thể sử dụng tốc độ 12-15 khung hình/giây vào những thời điểm nhất định. Người đọc môi có kinh nghiệm và người sử dụng ngôn ngữ ký hiệu cũng có lợi thế là đoán từ dựa vào kinh nghiệm. Như vậy, một số người dùng có thể có các cuộc hội thoại ngắn trên các kết nối chất lượng thấp hơn so với những yêu cầu chỉ ra ở trên.

5.1.5 Tính chất của phân giải thời gian

Trong hầu hết trường hợp, một máy quay được sử dụng cho truyền thông hình ảnh tuân theo các tiêu chuẩn hình ảnh nói chung, nghĩa là cung cấp tốc độ 25 hoặc 30 khung hình/giây. Trong cách sử dụng máy quay như vậy, không có nhiều điểm nổi bật khi xét đến tốc độ khung hình từ 12,5 đến 25 khung hình/giây hoặc từ 15 đến 30 khung hình/giây. Với khoảng tốc độ khung hình như vậy thì khoảng hình ảnh nguồn sẽ thay đổi tương ứng giữa 40 và 80 ms hoặc giữa 33 và 66 ms, gây ra nguy cơ thiếu các chi tiết chuyển động nhất định. Như vậy, để đáp ứng các yêu cầu cho 20 khung hình/giây với các máy quay thông thường, tốc độ khung hình mục tiêu nên là 25 hoặc 30 khung hình/giây.

5.2 Các yêu cầu độ phân giải không gian

Đối với phân giải không gian của các cuộc hội thoại ngôn ngữ ký hiệu giữa người và người cần lưu ý những đặc điểm sau đây:

- Có thể sử dụng phân giải QCIF nhưng các chi tiết nhỏ nhất biểu diễn hướng mắt nhìn bị mất. Điều này gây căng thẳng cho người nhận;

- Phân giải CIF rất thích hợp. Việc tăng từ QCIF tới CIF cho nhận thức ngôn ngữ tốt hơn;

- Phân giải SQCIF không thích hợp cho nhận thức tin cậy, hiếm khi ký hiệu được cảm nhận chính xác;

- Nếu độ phân giải khác nhau được sử dụng cho các phần khác nhau của hình ảnh, bàn tay và khuôn mặt sẽ yêu cầu độ phân giải cao nhất. Khi đó, cần lưu ý để không gây ra méo trong các phần khác của bức ảnh làm cho người sử dụng phân tâm.

Khung hình hiển thị trong trao đổi ngôn ngữ ký hiệu thường được để là 1 nửa người tính từ đầu đến bụng, ngón tay chiếm khoảng 1/50 bề rộng của hình ảnh. Để phân tích hình ảnh các ngón tay chính xác, một ngón tay được biểu diễn bởi ít nhất là 3 pixel. Điều đó đặt ra yêu cầu độ phân giải không gian tối thiểu là QCIF, có chứa 176 điểm ảnh rộng. Hướng mắt nhìn cũng rất quan trọng trong ngôn ngữ thị giác và đòi hỏi độ phân giải cao hơn. Vì vậy phân giải CIF là lựa chọn thích hợp.

Đối với đọc môi, phạm vi quan sát các cuộc hội thoại được giảm xuống thấp hơn đầu một chút. Trong trường hợp này, QCIF được xem là có độ phân giải đầy đủ cho đọc môi. Khi sử dụng độ phân giải QCIF, người sử dụng đầu cuối phải chắc chắn rằng màn hình hiển thị được xem ở khoảng cách thích hợp để độ phân giải tương đối thấp không gây thêm sự khó khăn cho nhận thức.

5.3 Độ chính xác

Trong truyền thông hình ảnh, bóng hình (độ mờ) xuất hiện khi có các chuyển động.

Các mô hình để mô tả bóng hình (độ mờ) cũng không phát triển rộng rãi. Có nhiều loại bóng hình khác nhau và gây ảnh hưởng khác nhau đến nhận nhận thức.

Hệ thống hình ảnh gia đình (VHS) có khả năng nhận thức tốt về ngôn ngữ ký hiệu và đọc môi. Trong ghi hình, các đối tượng di chuyển nhanh thường được hiển thị với độ mờ đáng kể bởi vì tốc độ màn trập thường là 1/50 đến 1/60 giây. Điều này cho thấy rằng độ mờ có thể chấp nhận được đối với các đối tượng di chuyển nhanh liên quan đến các cử động phức tạp trong ngôn ngữ ký hiệu.

Trong trường hợp cử động phức tạp, một số độ mờ đôi khi có thể xuất hiện. Độ phân giải không gian trong các cử động đó không được thấp hơn SQCIF.

Để nhận thức tốt, khi CIF là độ phân giải không gian cơ bản, độ mờ xuất hiện không nên vượt quá những gì được cảm nhận ở độ phân giải QCIF.

5.4 Độ trễ

Trễ hình ảnh đầu cuối đến đầu cuối, từ máy quay gửi đi đến thiết bị hiển thị nhận, được chuẩn hóa trong ứng dụng thoại. Các giá trị thích hợp là nhỏ hơn 0,4 s và có thể giảm xuống 0,1 s.

Các giá trị lớn hơn 0,8 s sẽ làm cản trở chất lượng của cuộc nói chuyện bằng ký hiệu.

Các yêu cầu đối với ngôn ngữ ký hiệu và đọc môi cũng tương tự với yêu cầu hội thoại. Thời gian từ khi phát âm cho đến khi đáp ứng mong đợi được nhìn thấy hoặc nghe thấy, có ít nhất hai lần bị trễ. Do đó, với giới hạn 0,4 s theo quy định của Khuyến nghị G.114 thì trễ một đáp ứng là 0,8 s.

5.5 Tính đồng bộ

Đối với thoại được hỗ trợ bởi đọc môi, tính đồng bộ giữa âm thanh và hình ảnh rất cần thiết. Sự sai khác thời gian lên tới 100 ms là có thể chấp nhận được.

Đối với những người sử dụng cả thoại và đọc môi, sự kết hợp đó rất hiệu quả cho nhận thức.

5.6 Các yêu cầu hiệu năng

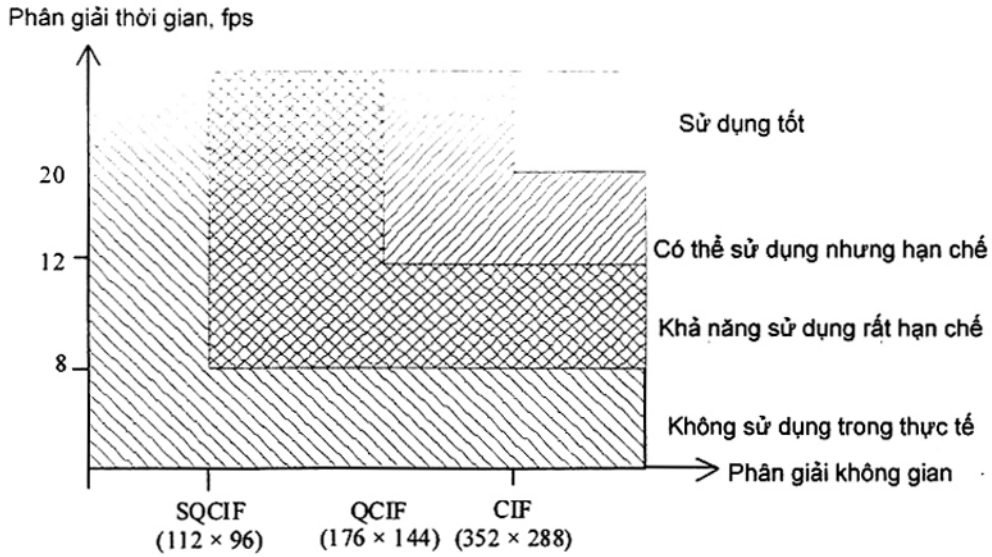
Đối với ứng dụng truyền tải ngôn ngữ ký hiệu và đọc môi trong cuộc hội thoại giữa người - người, các vấn đề hiệu năng cơ bản sau đây cần được áp dụng:

- Sử dụng tốc độ 25-30 khung hình/giây tại độ phân giải không gian CIF và độ trễ tối đa 0,4 s, chấp nhận bóng hình ít hơn tương ứng với QCIF trong quá trình chuyển động trung bình;

- Chấp nhận tốc độ 12-15 khung hình/giây QCIF với chuyển động trung bình và sự suy giảm không thường xuyên tương ứng với SQCIF trong quá trình chuyển động ngôn ngữ ký hiệu phức tạp (trong môi trường tỷ lệ bit thấp);

- Giữ tính đồng bộ âm thanh tốt hơn 100 ms;

- Trễ đầu cuối tới đầu cuối nên nhỏ hơn 0,4 s. Trong trường hợp không thể tránh khỏi, chấp nhận lên tới 0,8 s.



Hình 1 - Các yêu cầu độ phân giải cho ngôn ngữ ký hiệu và đọc môi trong cuộc hội thoại người với người.

Bảng 1 - Tóm tắt sự suy giảm tính khả dụng gây ra bởi trễ và bóng hình

Tính khả dụng	Trễ đầu cuối tới đầu cuối	Bóng hình đôi khi xuất hiện trong quá trình chuyển động lớn	
		Với phân giải CIF	Với phân giải QCIF
Tốt	<0.4 s	Không	–
Có thể sử dụng với một số hạn chế	0.4 – 0.8 s	Giảm xuống ≅ QCIF	Không
Tính khả dụng bị giới hạn	0.8 – 1.2 s	Giảm xuống ≅ SQCIF	Giảm xuống ≅ SQCIF
Không có tính khả dụng trong thực tế	>1.2 s	Giảm xuống < SQCIF	Giảm xuống < SQCIF

6 Khuyến nghị cho thiết bị đầu cuối

Để đáp ứng yêu cầu người sử dụng, các tính năng sau phải được thực hiện tại đầu cuối:

- Thiết bị đầu cuối cần có một giao diện để kích hoạt hệ thống cảnh báo bên ngoài, ví dụ như đèn flash, bộ giao động nhỏ (bộ giao động bỏ túi), bộ giao động xem kích thước hoặc máy tạo âm thanh mạnh;

- Người dùng đôi khi cần phải trở lại cuộc hội thoại văn bản. Do đó, khuyến khích thực hiện các giao thức hội thoại văn bản T.140 ở đầu cuối;

- Ưu tiên các cuộc gọi tốc độ hơn 20 khung hình/giây và trễ dưới 0,4 s, sử dụng thuật toán không bỏ qua khung nào. Tốc độ khung hình cao tự động đưa ra cơ hội để đạt được trễ hợp lý;

- Độ lệch từ tất cả các yêu cầu chất lượng có thể được chấp nhận lên đến 2 s sau một dịch chuyển cảnh.

7 Khuyến nghị đối với người sử dụng

Người sử dụng nên chuẩn bị để sử dụng thiết bị trong một môi trường có điều kiện ánh sáng tốt và một khung cảnh nền rõ nét.

Phụ lục A

(Tham khảo)

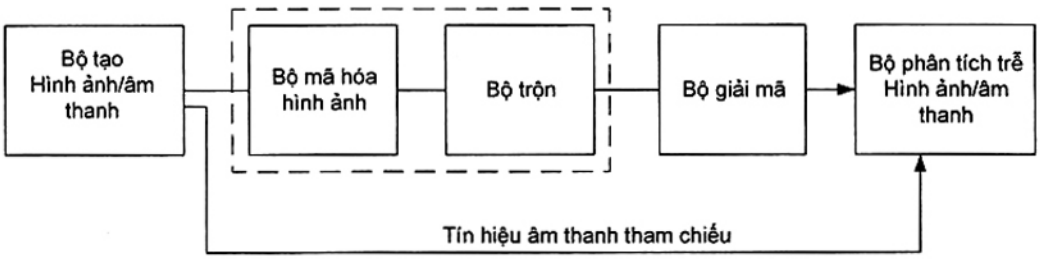
Phép đo

A.1 Đo độ trễ

Trễ tổng được xác định đơn giản bằng cách đo thời gian giữa lần kích hoạt được tạo bởi bộ tách sóng đặt ở đầu vào hệ thống và lần kích hoạt được tạo bởi bộ tách sóng tại đầu ra của bộ giải mã. Trễ tổng có thể được đo cho cả hình ảnh và âm thanh tùy thuộc vào đặc tính của bộ tách sóng. Độ chính xác của phép đo này là ± 1 ms.

Một phương pháp khác sử dụng đường âm thanh có sẵn như là một tín hiệu tham chiếu.

Thủ tục này dựa trên việc sử dụng thiết bị hiện có và hoạt động với một chuỗi định thời kiểm tra âm thanh và hình ảnh đặc biệt. Nó bao gồm một tông kiểm tra âm thanh và tín hiệu hình ảnh. Tông âm thanh bao gồm một sóng hình sin có tần số được chọn trong khoảng 1-10 kHz và các mức được chọn từ -20 dBu đến +20 dBu. Tín hiệu hình ảnh bao gồm một quá trình chuyển đổi độ sáng đen sang trắng ở dòng 45 cho các khuôn dạng 525 dòng và dòng 38 cho khuôn dạng 625 dòng.



Hình A.1 - Sơ đồ bài đo trễ hình ảnh

Thiết lập bài đo trễ hình ảnh tổng được mô tả trên Hình A.1. Lưu ý rằng tín hiệu âm thanh được cấp trực tiếp cho bài đo như một tham chiếu định thời.

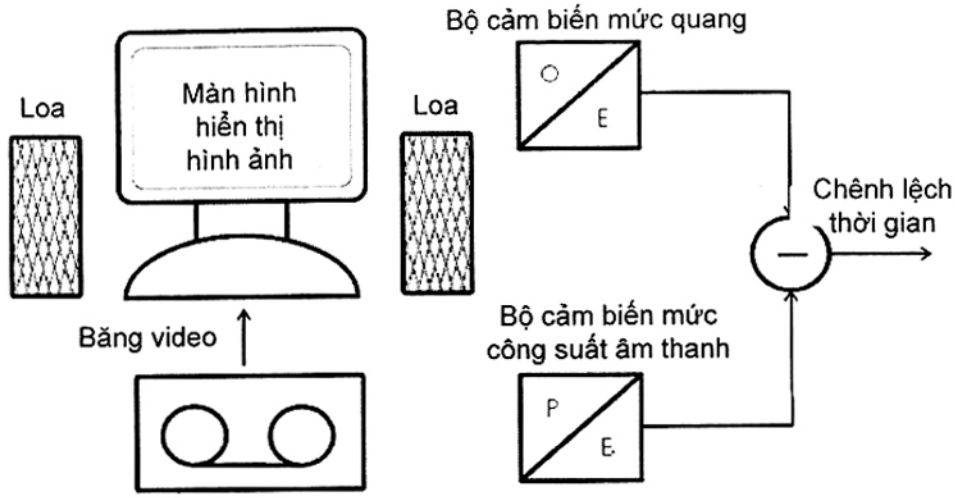
A.2 Đo đồng bộ thời gian giữa hình ảnh và âm thanh

Trong Hình A.2, băng video dùng để tham chiếu thường ở dưới dạng tệp được ghi trong môi trường như đĩa quang-từ và bộ nhớ lớn sẽ được tái tạo bằng thiết bị biểu diễn âm thanh – hình ảnh.

Những thay đổi tức thời ánh sáng đầu ra trung bình từ màn hình hiển thị màu sắc sẽ được cảm nhận bởi một bộ chuyển đổi quang điện tử để thu được những thay đổi tức thời tương ứng như tín hiệu điện. Đây là sự mô phỏng con mắt của người xem. Mạch bộ theo dõi phát của các tranzito quang điện được sử dụng cho mục đích này.

Những thay đổi tức thời mức áp suất âm thanh trung bình sẽ được cảm nhận bởi một bộ cảm biến mức áp suất âm thanh như một microphone đơn giản để thu được những thay đổi tức thời tương ứng như tín hiệu điện. Đây là sự mô phỏng tai của người nghe.

Cả tín hiệu âm thanh và hình ảnh được so sánh trong miền thời gian và sự khác nhau giữa các tín hiệu được đo sử dụng máy dao động dưới dạng miligiây.



Hình A.2 - Thiết lập bài đo chênh lệch thời gian giữa hình và tiếng

Phụ lục B

(Tham khảo)

Phương pháp đo chất lượng Video

Một codec hoặc một thiết bị đầu cuối, được kiểm tra bằng cách truyền những cảnh để đánh giá thông qua codec hoặc thông qua một tập hợp các điện thoại thấy hình được kết nối mạng. Kết quả kiểm tra được ghi nhận và đánh giá. Khuyến nghị P.931 quy định cụ thể một phương pháp đánh giá.

B.1 Phương pháp đo lỗi bình phương trung bình

Mục B.1 cung cấp các phương pháp đo cho một hệ thống sử dụng phương pháp lỗi bình phương trung bình. Triển khai phương pháp này có thể cung cấp chuỗi khung hình video thích hợp tại đầu vào kênh. Phương pháp này cũng đòi hỏi phải bắt giữ và nếu cần thiết số hóa thành phần độ sáng của chuỗi khung hình video tại các giao diện kênh.

B.1.1 Khái quát chung

Phát hiện các khung tích cực trong chuỗi các khung video và tìm kiếm các khung ánh xạ thích hợp giữa các chuỗi đòi hỏi một phương pháp so sánh chuẩn. Phương pháp này so sánh các khung video trên cơ sở pixel-by-pixel và tóm lược sự khác biệt giữa một cặp khung như lỗi bình phương trung bình trên tất cả các điểm ảnh được quan tâm. Vì vậy, đối với một cặp khung (một từ chuỗi đầu vào và một từ chuỗi đầu ra) Lỗi bình phương trung bình (MSE) được tính theo công thức:

$$M\left[V'(m),V(n)\right]=\frac{1}{K_s}\sum_{j=J_{min}}^{J_{max}}\sum_{i=I_{min}}^{I_{max}}\left[V'(i,j,m)-V(i,j,n)\right]^2$$

trong đó V'(i,j,m) là giá trị điểm ảnh i,j trong khung đầu ra tại thời điểm T'(m) và V(i,j,n) là giá trị điểm ảnh i,j trong khung đầu vào tại thời điểm T(n). K_s là tổng số điểm ảnh quan tâm trong phân khung chữ nhật, được cho bởi công thức:

$$K_s = (I_{max} - I_{min} + 1) \times (J_{max} - J_{min} + 1)$$

Lưu ý V'(i,j,m) được điều chỉnh với bất kỳ độ lợi, độ lệch mức, dịch chuyển ngang, dịch chuyển dọc và lấy tỷ lệ không gian (nếu cần) giữa đầu vào và đầu ra (với các hệ số điều chỉnh tương ứng g, l, h, v và z):

$$V'(i,j,m)=\frac{V^*(x+v,y+h,m)-1}{g}$$

trong đó V*(i,j,m) là điểm ảnh đầu ra trước khi áp dụng các hệ số điều chỉnh. Nếu video đầu ra được định cỡ lại phù hợp với đầu vào, thì:

$$V'(i,j,m)=\frac{V^{**}(i+v,j+h,m)-1}{g}$$

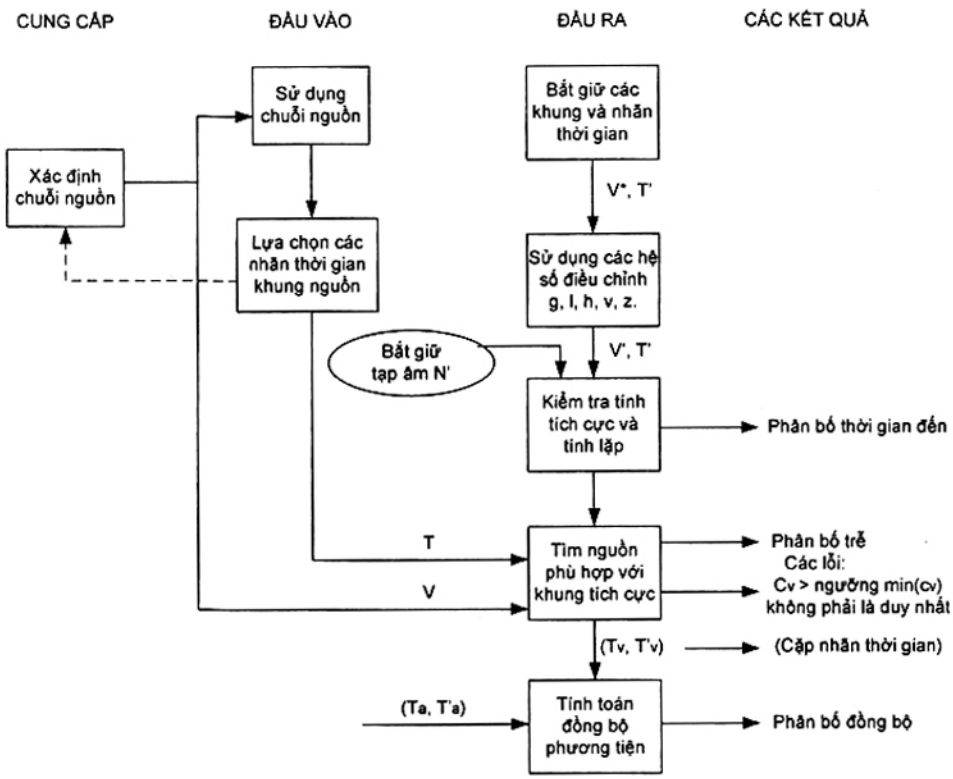
trong đó V**(m) = f(V*(m), z)

và f(V*(m), z) biểu diễn hàm định cỡ lại

Để so sánh giữa các khung gần kề trong một chuỗi (ví dụ như tìm các khung tích cực tại giao diện đầu ra), V(i,j,n) trở thành V'(i,j,m-1) trong phương trình MSE ở trên.

MSE là một hệ số quan trọng để tính tỉ số tín hiệu trên tạp âm đỉnh (PSNR):

$$PSNR = 20\log_{10}\left[\frac{V_{peak}}{\sqrt{M[V'(m),V(n)]}}\right] \text{dB}$$



Hình B.1 - Sơ đồ thuật toán để đo chất lượng video dựa vào MSE

B.1.2 Xác định sự khác biệt tối thiểu có thể phân biệt giữa các khung hình

Phần này đặc tả các phương pháp xác định tạp âm (hoặc phương sai không mong muốn) trong các quá trình số hóa và lưu trữ để lựa chọn chuỗi khung video cho so sánh. Mức tạp âm này phụ thuộc vào quá trình lựa chọn cụ thể (ví dụ như khuôn dạng số hóa) và được biết trước để thực hiện các phép đo có giá trị.

Các điều kiện kiểm tra để hiệu chỉnh tạp âm bắt được như sau:

- Sử dụng một cảnh video tĩnh cho đầu vào kênh. Video tĩnh được định nghĩa là "hình ảnh video không có chuyển động hoặc thay đổi". Điều quan trọng là duy trì tỉ số tín hiệu trên tạp âm video đầu vào giống nhau trong quá trình hiệu chỉnh và đo lường. Một kỹ thuật sử dụng chuỗi nguồn bao gồm một khung

được lập từ một hoặc nhiều đoạn video chuyển động dành cho kiểm tra sau này. Kỹ thuật này không tái tạo tạp âm trong chuỗi nguồn và chỉ thích hợp với trường hợp tạp âm bắt được cao hơn đáng kể so với tạp âm nguồn. Đối với một số chuỗi kiểm tra và video trực tiếp, có thể chia khung video theo không gian và xác định một phân khung tĩnh (ví dụ như nền) cho hiệu chỉnh và một phân khung chuyển động cho các phép đo khác. Các tín hiệu kiểm tra tĩnh (các thanh màu SMPTE) được sử dụng hiệu quả (tạp âm nguồn có mặt trong tín hiệu kiểm tra tĩnh giống như trong chuỗi kiểm tra);

- Bắt giữ (số hóa và lưu trữ) chuỗi các khung tương ứng tại đầu ra kênh. Khoảng 30-60 khung là đủ. Khi kênh sử dụng nén kỹ thuật số thì kênh được phép đạt đến một mức chất lượng ổn định, do đó tránh bất kỳ đáp ứng cắt cảnh nào làm sai lệch phép đo tạp âm.

Nhìn chung, tạp âm bắt ở đầu vào sẽ khác với tạp âm ở đầu ra. Một số codec lọc ra tạp âm nguồn để cải thiện tín hiệu cho mã hóa.

Với một chuỗi 30 khung, tính tập 30 -1 =29 giá trị MSE khung liền kề , M[V'(m), V'(m – 1)].

Mức tạp âm bắt đầu ra là giá trị MSE cực đại của tập

$$v'(\mathrm{m}) = \mathrm{M}[\mathrm{V}'(\mathrm{m}).\ \mathrm{V}'(\mathrm{m}-1)] \text{ for } \mathrm{m} = 2,\ 3,\ \ldots\ 30$$

trong đó $v'(m)$ là giá trị MSE cho khung V'(m) và tạp âm bắt giữ N' là:

$$N' = \max(v')$$

trong đó v' là tập giá trị MSE cho chuỗi V'. Mức độ biến thiên trong tập các giá trị MSE nên nhỏ hơn 20% do tính trung bình nhiều điểm ảnh cho mỗi giá trị trong tập. Đối với các chuỗi đầu vào, ta có N = max(v).

Cho phép một số dung sai giữa mức tạp âm bắt giữ và ngưỡng để phát hiện các khung tích cực, định nghĩa khung đầu ra mà có $v'(m)$) = M [V'(m), V' (m - 1)] ≤ 1,5 x N' là các khung lặp. Đối với một chuỗi mã nguồn, định nghĩa khung mà $v'(n)$ = M [V (n), V (n - 1)] ≤ 1,5 x N là các khung đồng nhất. Có thể có sự khác biệt nhỏ giữa các khung lặp hoặc các khung đồng nhất, tuy nhiên hệ thống đo lường không thể phát hiện chính xác được. Việc lựa chọn và xác định chuỗi nguồn để kiểm tra sẽ đưa ngưỡng này vào bản kê. Không thể phát hiện các khung tích cực khi các khung trong chuỗi nguồn đồng nhất cho các thiết bị đo lường. Dung sai này cũng thúc đẩy việc phát hiện khung tích cực với độ tin cậy lớn hơn.

B.1.3 Kiểm tra chuỗi với sự khác biệt rõ ràng

Đối với một chuỗi video V, tính tập các giá trị MSE v và so sánh mỗi giá trị của tập với ngưỡng của các khung đồng nhất (1,5 x N). Tất cả khung V (n) thỏa mãn M [V (n), V (n - 1)]> 1,5 × N có sự khác biệt phân biệt được từ khung trước đó. Một kênh thử nghiệm sẽ được cung cấp các khung đầu vào có sự khác biệt rõ ràng để kiểm tra các khung tích cực và các khung lặp.

Khi xem xét chuỗi nguồn sử dụng cho hệ thống truyền dẫn chất lượng cao mà bảo toàn tính toàn vẹn trường, việc so sánh khung Video hiện tại V (n) với V (n - 2) rất thích hợp để ghép các trường tương đương và tránh lỗi so sánh từ độ lệch không gian giữa các trường.

Thủ tục cho các bài kiểm tra điều kiện sau đây đảm bảo cho một khung Video có khác biệt rõ rệt (với chuỗi nguồn có các trường xen kẽ):

- Tính M[V(n), V(n - 1)];
- Nếu kết quả ≤ 1,5N, ghi không thể phân biệt được các khung, ngược lại sẽ tiếp tục;
- Tính M[V(n), V(n - 2)];
- Nếu kết quả ≤ 1,5N, ghi không thể phân biệt được các khung, ngược lại sẽ tiếp tục;
- Khung V (n) có sự khác biệt có khả năng phân biệt được.

B.1.4 Phân loại khung tích cực và khung lặp

Đối với chuỗi video đầu ra V', tính tập giá trị MSE M [V'(m), V'(m - 1)] và so sánh mỗi giá trị trong tập với ngưỡng cho khung đồng nhất (1,5 x N').

Lưu ý rằng nhiều hệ thống truyền dẫn chất lượng cao bảo đảm tính toàn vẹn trường, đồng thời cũng sinh ra méo tối thiểu. Các hệ thống này cũng thích hợp để so sánh khung Video hiện tại V'(m) với V'(m - 2) để ghép các trường tương đương và tránh lỗi so sánh từ độ lệch không gian giữa các trường. Khi kiểm tra tại các giao diện không đan kẽ hoặc sử dụng các tùy chọn được công nhận để giảm tỷ lệ bắt giữ và độ phân giải thì việc so sánh với V' (m - 2) là không cần thiết.

Một khung V' (m) mà MSE có kết quả là M [V' (m), V' (m - 1)] và M [V' (m), V' (m - 2)]> 1,5 x N', đáp ứng chuỗi đầu vào có sự khác biệt rõ ràng, được giới hạn tương ứng với mỗi khung và được phân loại là khung tích cực.

Một khung V' (m) mà MSE có kết quả là M [V' (m), V' (m - 1)] và M [V' (m), V' (m - 2)]≤ 1,5 x N', đáp ứng chuỗi đầu vào có sự khác biệt rõ ràng, tương ứng với V' (m - 1) hoặc V' (m - 2) thì được phân loại là khung lặp.

B.1.5 Kiểm tra sự tương ứng giữa các khung (các khung thích hợp)

Đối với khung tích cực m và chuỗi đầu vào khung X, tính tập X giá trị MSE, M [V' (m), V]. Khung đầu vào với sự tương ứng tốt nhất là khung mà tạo ra giá trị MSE tối thiểu trong tập:

$$c_v(x) = M[V'(m), V(x)] \text{ với } 1 \le X \le X$$

c_v là tập giá trị MSE cho khung V' (m) so với mỗi khung trong chuỗi V và khung đầu vào phù hợp nhất V' (m) được định nghĩa là:

$$C_v = \min (c_v)$$

[Lỗi cực tiểu (MSE) biểu diễn sự tương ứng cực đại hoặc phù hợp nhất giữa các khung]

Một tập các quy tắc có thể cải thiện quá trình kết hợp và làm giảm sự không rõ ràng. Có thể có trường hợp mà một khung tích cực tương ứng chặt chẽ với nhiều hơn một khung đầu vào. Trường hợp này cần được giảm thiểu với tiêu chí phù hợp dựa trên phương pháp so sánh điểm ảnh (MSE), tuy nhiên một số trường hợp lại tăng khả năng không rõ ràng. Đó là:

- Méo không gian do tốc độ truyền bit thấp, sử dụng định dạng khung số độ phân giải thấp, vv...

- Nội dung nguồn - chuyển động nhiều (gây nhòe hay méo), chuyển động lặp lại, các khoảng tĩnh trong một chuỗi;
- Tốc độ khung đầu ra tích cực thấp cho phép nhiều khung nguồn phù hợp nhất có thể;
- Sử dụng nội suy khung khiến cho quá trình phù hợp khó khăn hơn.

Các quy tắc sau đây rất hữu ích để giải quyết sự phù hợp không rõ ràng:

- Yêu cầu phù hợp một - một: Chỉ có một khung tích cực phù hợp với một khung đầu vào cho trước. Do đó có thể giải thích phù hợp kép là một khung tích cực được phát hiện sai. Nếu kết quả so sánh có phù hợp kép thì tình trạng lỗi này phải được báo cáo;
- Chuỗi bắt buộc: Ví dụ (với các khung không đan xen) nếu V' (m) phù hợp với V (n), thì các khung tích cực tiếp theo V' (m + 2) phải phù hợp với V (n + 1) hoặc V (n + 2) hoặc V (n + 3), vv. V' (m + 2) không được phép phù hợp với V (n - 1) hoặc V (n);
- Trễ tối thiểu được chấp nhận: trễ tối thiểu là tmin ≥ 0;
- Nhận dạng điều kiện không thích hợp: Một số khung tích cực có méo quá nhiều để phù hợp với chuỗi truyền. Các khung như vậy sẽ được tính và báo cáo, cùng với ngưỡng không phù hợp được sử dụng. Người dùng hệ thống đo phải xác định phạm vi thông dụng của các giá trị MSE thích hợp cho hệ thống truyền tải thử nghiệm, và thiết lập ngưỡng trên phạm vi này;
- Chẩn đoán: Quá trình kết hợp có thể được lặp lại từ đầu ngược lại của chuỗi để xem xét nếu sự phù hợp không rõ ràng ít hơn và xảy ra sự không phù hợp. Chiều hướng của quy định này cũng phải được đảo ngược;
- Kiểm tra khung tiếp theo: Nếu khung tích cực tiếp theo trong chuỗi đầu ra có sự phù hợp duy nhất trong chuỗi truyền, thì sử dụng sự phù hợp của nó và thực thi các quy tắc ở trên vào khung tích cực trước đó;
- Lựa chọn ngẫu nhiên: Khi sự không rõ ràng vẫn tồn tại, lựa chọn ngẫu nhiên có thể được sử dụng. Tuy nhiên, khuyến nghị tính toán MSE với độ phân giải đủ để giảm thiểu tình trạng như vậy. Lỗi sinh ra phân phối bởi quá trình ngẫu nhiên được loại bỏ trên một chuỗi và các thống kê đặc trưng không bị ảnh hưởng. Những lựa chọn như vậy sẽ được tính và báo cáo;
- Nếu các kết quả sử dụng một cảnh cụ thể có xu hướng yêu cầu độ phân giải và can thiệp mở rộng bằng cách sử dụng những quy tắc này thì phép đo nên sử dụng một cảnh khác.

B.1.6 Xác định chuỗi nguồn cho phương pháp lỗi bình phương trung bình

Sự thành công của các phương pháp dựa vào MSE phụ thuộc vào việc sử dụng các chuỗi nguồn thích hợp. Như đã nêu ở trên, chuỗi nguồn sẽ khác nhau giữa khung với khung mà có thể phân biệt bằng các thiết bị đo lường, tránh chuyển động lặp lại, và tránh những khoảng video tĩnh gây ra sự phù hợp không rõ ràng. Khi sử dụng một vùng phân khung, quá trình xử lý sẽ thông qua một vùng tương tự là cơ sở để xác định. Các thủ tục sau đây có thể xác định các đoạn hình ảnh phù hợp cho đo lường:

- Lấy khung video đầu tiên trong chuỗi nguồn và so sánh nó với tất cả các khung khác trong chuỗi;
- Ghi lại số lượng và vị trí của tất cả các khung đồng nhất (như mô tả trong C.1.2);

- Phân tích: Khoảng thời gian giữa các khung đồng nhất đủ để giải quyết sự không rõ ràng liên kết đầu vào-đầu ra bằng cách sử dụng ước tính ưu tiên khoảng thời gian đến khung và các thông tin khác. Ví dụ, nếu biết trễ truyền tải < 2s thì thời gian các khung đồng nhất có thể xuất hiện riêng ≥ 2s;
- Lặp lại các bước trên ít nhất là với X khung đầu tiên (ví dụ: X = 60). Tập trung vào các khung đầu trong chuỗi vì các kết quả so sánh phụ thuộc vào các kết quả trước.

B.1.7 Các hệ số ảnh hưởng đến tính ổn định và độ chính xác đo lường

Trong nhiều hệ thống truyền tải video, bộ giải mã phải cung cấp các khung video tại đầu ra của nó theo một chế độ hiển thị định kỳ (chẳng hạn như giao diện phức hợp tương tự). Nếu đồng hồ hiển thị đầu vào và đầu ra không đồng bộ thì phải thêm bộ đệm vào bộ giải mã. Khi bộ giải mã có một khung Video đã sẵn sàng cho hiển thị, nó vẫn phải chờ đến cơ hội đầu ra tiếp theo và do đó làm tăng trễ của toàn hệ thống. Gọi khoảng thời gian đợi giải mã là trễ đầu ra.

Trễ đầu ra được giới hạn bởi khoảng thời gian giữa các bản cập nhật hiển thị. Đối với hệ thống truyền dẫn có giao diện phức hợp mà có thể cập nhật các giới hạn trường thì trễ đầu ra cực đại là 16,7 ms. Với cập nhật giới hạn trường 525 dòng thì trễ cực đại là 33 ms. Trễ đầu ra thực tế là các giá trị ngẫu nhiên từ 0 đến giá trị cực đại.

Khi đồng hồ dạng sóng đầu vào và đầu ra có khoảng lệch tần số nhỏ thì trễ đầu ra sẽ thay đổi theo thời gian. Nếu độ lệch tần số là không đổi thì trễ đầu ra sẽ thay đổi trên phạm vi của nó trong khoảng thời gian chính xác. Nếu các đồng hồ được đồng bộ hóa với các bộ dao động kiểm soát tia Cesium độc lập thì trễ đầu ra sẽ thay đổi <1 ms trong 13.900 giờ. Nếu độ chính xác đồng hồ được lấy từ bộ dao động thạch anh độc lập (với độc lệch trang 2ppm) thì trễ đầu ra xoay quanh dải 0 - 33 ms trong 4,58 giờ.

Nhãn thời gian có đặc tính cho phép phân giải trường con của trễ đầu ra là một thành phần chặt chẽ của trễ toàn hệ thống truyền tải.

B.2 Phương pháp đo mã hóa thời gian trong khung

Trong một số trường hợp, có thể chèn các ký hiệu hiện trong đoạn video đầu vào được sử dụng để xác định mỗi khung đầu vào. Các ký hiệu này được đưa tới đầu ra của hệ thống và được sử dụng để đo trễ và tốc độ khung hình.

Phụ lục C

(Tham khảo)

Chuỗi kiểm tra Irene

C.1 Đánh vần bằng tay

Bảng C.1 cho thấy một miêu tả gần đúng một chuỗi đánh vần bằng tay trong chuỗi thử nghiệm "Irene". Những hình ảnh từ chuỗi này được thể hiện trong Hình C.1.

Bảng C.1 - Ví dụ biểu diễn đánh vần bằng tay trong các khung hình tốc độ 25 và 12,5 khung trên giây

Số khung hình	308		310					315					320					325				330					335	336
25 fps	e	e	e	–	d	s	s	s	s	–	v	v	v	–	i	–	–	k	k	k	–	e	n	n	n	n	n	n	n
12.5 fps		e		–		s		s		–		v		–		–		k		k		e		n		n		n	

Các số trên hàng trên cùng là các số khung bắt đầu của chuỗi. Các chữ biểu thị khi các chữ cái khá rõ ràng được hình thành bởi bàn tay. Dấu gạch ngang chỉ ra rằng không có chữ cái nào được hình thành trong quá trình chuyển tiếp giữa các chữ cái. Từ này là "Edsviken", tên một địa danh.

Trong số 8 chữ cái này, ba chữ cái chỉ thấy trên một khung hình và do đó sẽ có nguy cơ bị mất tại 12,5 khung hình/giây. Ví dụ mẫu khung 12,5 fps được cho ở hàng dưới của bảng, trong đó có 2 chữ cái bị mất, từ gốc "Edsviken" chỉ còn lại là "Esvken" (xem Hình C.2). Điều này rõ ràng thể hiện nguy cơ mất nội dung ngôn ngữ khi tỷ lệ khung hình thấp hơn 20 fps.

Phân phối chữ cái trong chuỗi 25 fps:

1 khung 3 chữ cái;

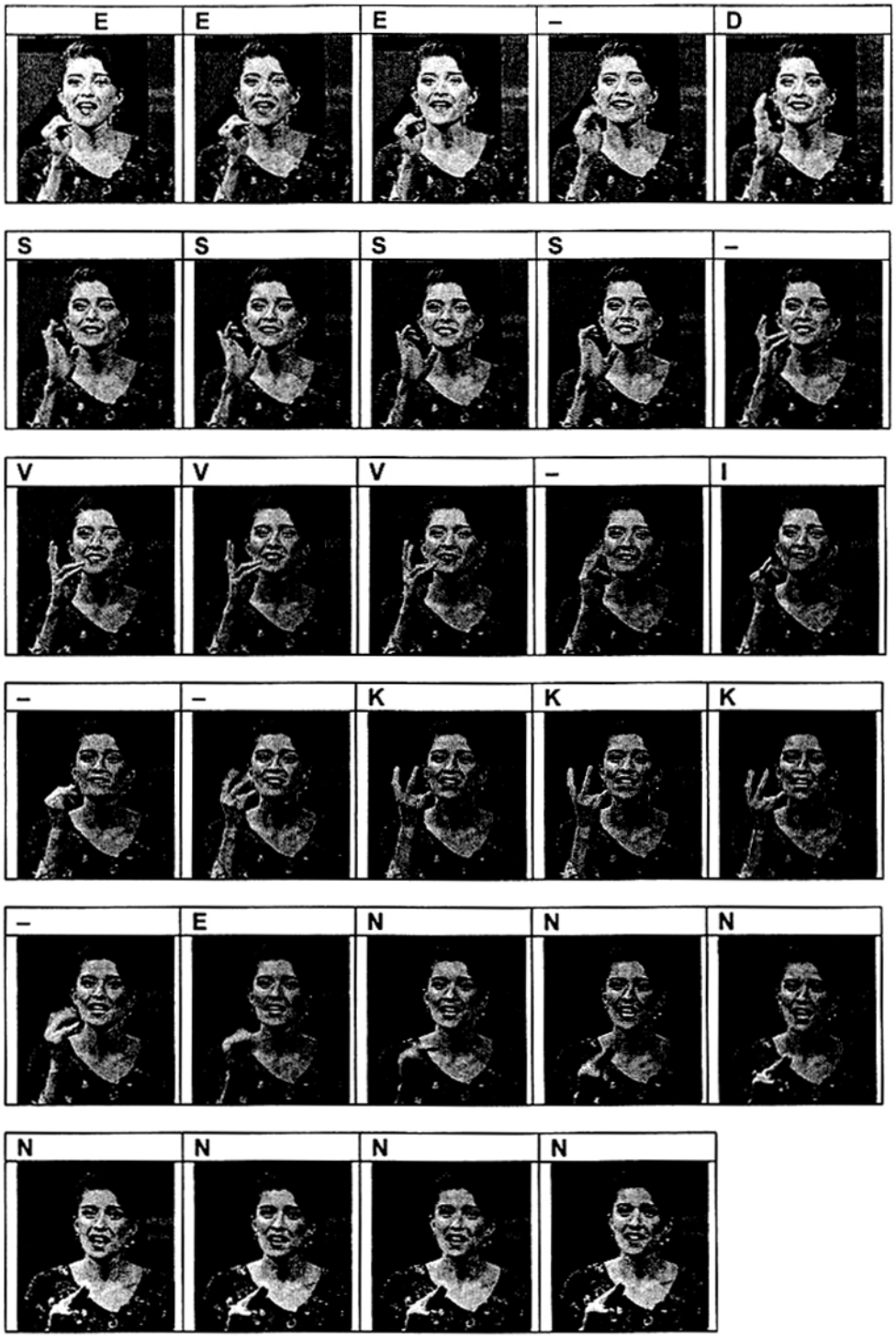
2 khung 0 chữ cái;

3 khung 3 chữ cái;

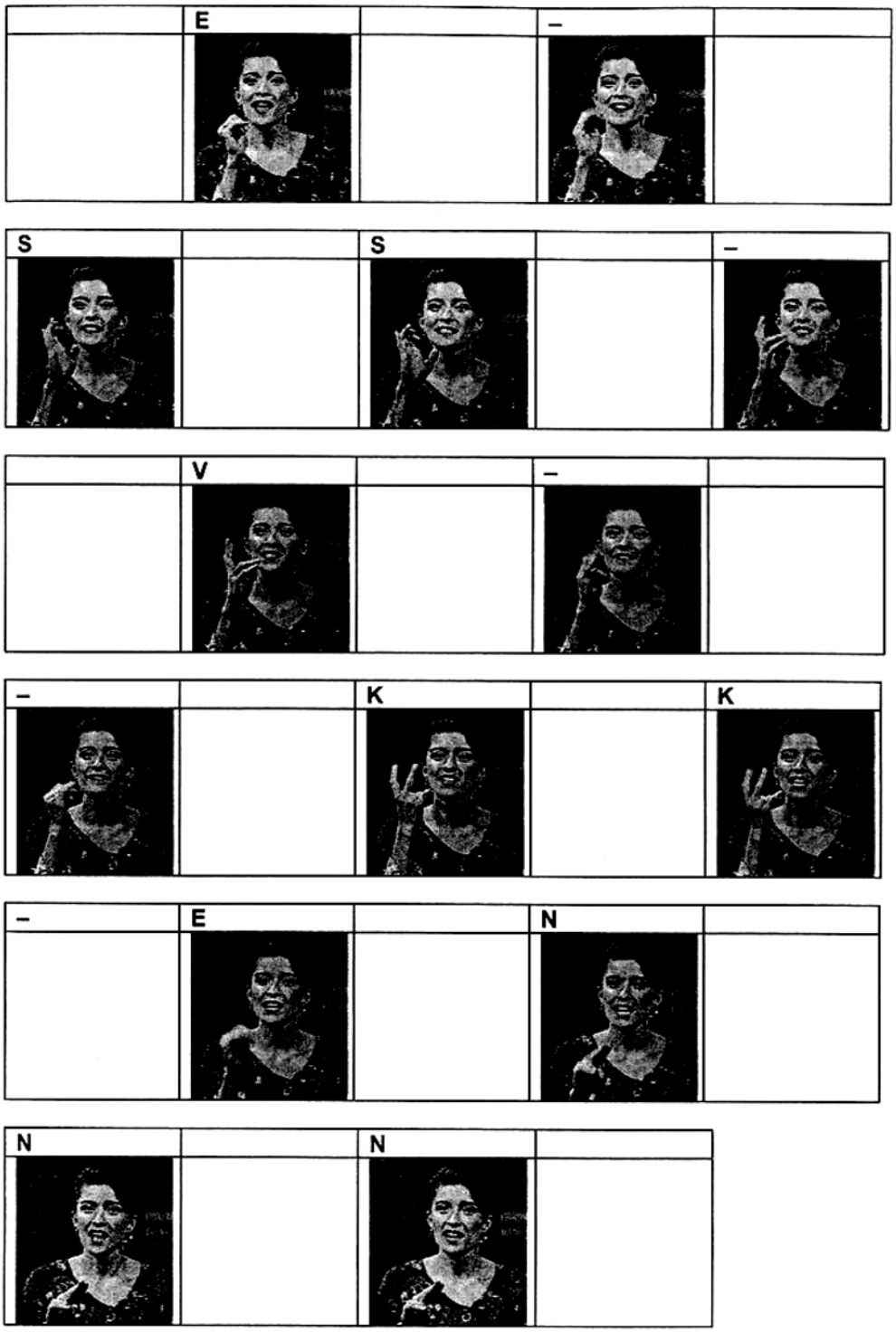
4 khung 1 chữ cái;

7 khung 1 chữ cái (cuối cụm).

Chiều dài trung bình trong cụm là: 2,3 khung / chữ cái.



Hình C.1 - Các khung hình chứa từ được đánh vần bằng tay "Edsviken" ghi tại 25fps



Hình C.2 - Các khung hình chứa từ được đánh vần bằng tay "Edsviken" ghi tại 12,5 fps. 2 chữ cái bị mất

Trong ví dụ này, các chữ cái trong từ thay đổi từ 1 đến 4 khung, với mỗi khung biểu diễn 40 ms. Chiều dài trung bình là 2,3 khung xuất hiện cho 1 chữ cái. Ví dụ trên chưa đủ để thực hiện bất kỳ kết luận thống kê thực tế. Tuy nhiên, có thể thấy rằng, với tốc độ đánh vần bằng tay, tốc độ khung hình 25 fps là đủ, trong khi 12,5 fps sẽ yêu cầu một số phỏng đoán để cảm nhận được các từ được đánh vần bằng tay.

C.2 Ký hiệu chung

Phần lớn đoạn phim "Irene" sử dụng các ký hiệu và không sử dụng đánh vần bằng tay.

Một phân tích đơn giản được thực hiện trên cụm từ sau. Nó được sao chép lại bởi các ký hiệu với số lượng khung cho mỗi ký hiệu trong dấu ngoặc đơn.

Chuỗi được biểu diễn giữa khung 406 và 520 trong chuỗi "Irene"

"SHE(7) TELLS(7) SELF(11) HOW(4) SHE(2) FELT(11) EXPERIENCED(13) ADOLESCENCE(16)."

Không có ký hiệu nào trong chuỗi ngắn hơn 2 khung và không có ký hiệu nào gồm nhiều cử động nhanh hơn đánh vần bằng tay. Một số ký hiệu bao gồm nhiều cử động và do đó cần phải có các yêu cầu khác về mã hóa video.

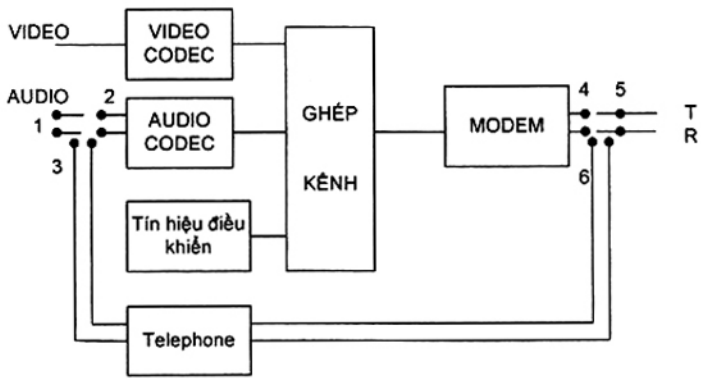
Phụ lục D

(Tham khảo)

Chế độ thoại và chế độ thoại có hình của máy điện thoại thấy hình

D.1. Cấu tạo của máy điện thoại thấy hình tốc độ thấp

Đặc tính cơ bản của loại máy này là nó có tính năng vừa là máy điện thoại thông thường vừa là máy điện thoại thấy hình. Bản chất của tín hiệu thoại ở hai chế độ làm việc này là hoàn toàn khác nhau. Các chế độ này được mô tả trên Hình D.1.



Hình D.1 - Cấu trúc của máy điện thoại thấy hình

D.2. Chế độ thoại không hình

Ở chế độ thoại không hình máy làm việc như một máy điện thoại thông thường. Khi đó các tiếp điểm 1-2 và 4-5 ở chế độ ngắt. Các tiếp điểm 1-3 và 5-6 được nối với nhau. Như vậy ở chế độ này máy hoàn toàn không cần đến các bộ mã hóa và giải mã cũng như các bộ ghép, tách kênh và mô- đem. Vì vậy máy chỉ làm việc với nguồn cấp qua hai dây thoại.

D.3 Chế độ thoại thấy hình

Ở chế độ thoại thấy hình, các tiếp điểm 1-3 và 4-6 ngắt còn các tiếp điểm 1-2 và 5-4 được nối với nhau. Ở chế độ này, tín hiệu thoại và tín hiệu hình thực chất đã được xử lý thành tín hiệu số..

- Tín hiệu hình:

Tín hiện hình được biến đoi trong quá trình xử lý hết sức phức tạp nhờ các phép mã hoá nội suy hình ảnh, các phép mã hoá côsin rời rạc, các phép mã hoá có độ dài từ mã thay đổi... Kết quả là tốc độ bit dành cho ảnh nằm trong khoảng vài kbit/s.

- Tín hiệu thoại:

Tín hiệu thoại được xử lý bằng phương pháp nén tiếng nói dùng kỹ thuật số. Kết quả là tín hiệu thoại được biến đổi thành luồng số tốc độ khoảng từ 6 đến 8 kbit/s.

- Ghép tín hiệu:

Tín hiệu hình và thoại đã qua xử lý được ghép cùng với tín hiệu điều khiển và được mã hóa thành một luồng tín hiệu. Luồng tín hiệu số này được đưa tới môđem để điều chế và truyền đi trên đường điện thoại.

Đặc điểm cơ bản của loại mô-đem này là thời gian bắt tay giữa hai máy rất ngắn. Thông thường vì chất lượng đường truyền khác nhau nên mô-đem được thiết kế với vài tốc độ khác nhau. Kết quả là tùy theo chất lượng đường truyền mà chất lượng hình và thoại sẽ khác nhau.

Trong tiêu chuẩn này có đưa ra các tiêu chuẩn cho hai chế độ khác nhau do tính chất hoàn toàn khác nhau của hai chế độ thoại không thấy hình và và thoại thấy hình.

Thư mục tài liệu tham khảo

[1] ITU-T H-series - Supplement 1 (05/1999), Sign language and lip reading real time conversation using low bit rate video communication (chất lượng dịch vụ video thoại tốc độ thấp sử dụng cho trao đổi ngôn ngữ ký hiệu và đọc môi thời gian thực).

[2] TCN68 - 154: 1995, Điện thoại thấy hình tốc độ thấp - Yêu cầu kỹ thuật.

[3] HELLSTRÖM, DELEVERT, REVELIUS: Quality requirements on Videotelephony for Sign Language, *Swedish National Association of the Deaf, 1997.* (Yêu cầu chất lượng dịch vụ video thoại sử dụng ngôn ngữ ký hiệu).

[4] ITU-T Recommendation G.114 (1996), *One-way transmission time.* (Thời gian truyền dẫn một chiều).

[5] FROWEIN: Improved speech reception through videotelephony, *IEEE journal on Selected Areas in Communication, May 1991.* (Cải tiến thu thoại qua điện thoại thấy hình)

[6] ITU-T P.931 (12/98) Multimedia communications delay, synchronization and frame rate measurement (Đo tốc độ khung, đồng bộ và trễ truyền thông đa phương tiện).

[7] IEC 100/AGS(Secretariat)216 (2006) Multimedia quality - Method of measurement and assessment of synchronization of audio and video.

[8] ETSI TR 101 290 V1.2.1 (2001-05) Digital Video Broadcasting (DVB); Measurement guidelines for DVB systems.