# RESIDUAL-GUIDED PERSONALIZED SPEECH SYNTHESIS BASED ON FACE IMAGE

*Jianrong Wang*[1], *Zixuan Wang*[1], *Xiaosheng Hu*[1], *Xuewei Li*[1], *Qiang Fang*[2], *Li Liu*[3,*]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
[3]Shenzhen Research Institute of Big Data, the Chinese University of Hong Kong, Shenzhen, China

## ABSTRACT

Previous works derive personalized speech features by training the model on a large dataset composed of his/her audio sounds. It was reported that face information has a strong link with the speech sound. Thus in this work, we innovatively extract personalized speech features from human faces to synthesize personalized speech using neural vocoder. A **F**ace-based **R**esidual **P**ersonalized **S**peech **S**ynthesis Model (FR-PSS) containing a speech encoder, a speech synthesizer and a face encoder is designed for PSS. In this model, by designing two speech priors, a residual-guided strategy is introduced to guide the face feature to approach the true speech feature in the training. Moreover, considering the error of feature's absolute values and their directional bias, we formulate a novel tri-item loss function for face encoder. Experimental results show that the speech synthesized by our model is comparable to the personalized speech synthesized by training a large amount of audio data in previous works.

***Index Terms***— Personalized speech synthesis, Speech prior, Residual, Attention mechanism

## 1. INTRODUCTION

Generating natural speech from text (text-to-speech synthesis, TTS) has been studied for decades [1, 2, 3]. Tacotron [1] and Tacotron 2 [2] are two efficient end-to-end speech synthesis models for TTS based on deep learning. As for the personalized speech synthesis of multiple speakers, Google proposed a system [4] that used several audio datasets to train a speaker encoder with a sequence-to-sequence TTS network and the vocoder based on Tacotron2. Besides, Baidu's deepvoice3 [5] added a speaker encoder on the basis of Tacotron to characterize the timbre characteristics of speakers.

Most of current works extracted personalized speech features based on a large volume of audio signals. Recently, some research works on speech to face generation using deep learning methods have emerged [6, 7, 8]. Motivated by these promising results, we believe that the inverse direction, *i.e.*, the personalized speech synthesis from one's face image is feasible. Face2Speech [9] is the method to predict speech features from face images and then synthesize speech using the WORLD vocoder [10]. However, it was reported in [5] that WORLD vocoder introduces various noticeable artifacts, and the WaveNet vocoder sounds more natural. Besides, [9] only performed qualitative experiments without quantitative experiments, thus lacking objective evaluation criteria.

In this work, we propose a novel speaker-independent *Face-based Residual Personalized Speech Synthesis Model (FR-PSS)* within an encoder-decoder architecture, which extracts personalized information from the face image, and synthesizes audio speech using natural vocoder. An overview of our proposed FR-PSS is shown in Fig. 1. Firstly, the speech encoder is trained to extract features from speech. Secondly, the speech synthesizer is trained to synthesize speech from a given text and features generated from the pre-trained speech encoder. Thirdly, the face encoder is trained with the pairs of face image and speech for making features extracted from a face image of a speaker closer to one derived from his/her speech. The tri-item loss function and a residual-guided strategy is introduced in this step. Finally, the face encoder and speech synthesizer are concatenated for building our FR-PSS model. Compared with Face2Speech, we conduct a series of quantitative experiments. Experimental results show that our FR-PSS achieves a comparable performance compared with method that synthesizes personalized speech using a speech-derived embedding vector.

Overall, our contributions can be summarized as follows. i) A new speech synthesis model FR-PSS is proposed. We extract personalized speech features from human faces to synthesize personalized speech using neural vocoder. ii) A residual-guided strategy is designed by incorporating a prior speech feature to make the network capture more representative face features and improve model learning efficiency. iii) We innovatively establish a tri-item loss function to accelerate training convergence.

## 2. FACE-BASED RESIDUAL PERSONALIZED SPEECH SYNTHESIS MODEL

In this section, we will introduce the proposed FR-PSS (shown in Fig. 1), which includes Speech Encoder (*SE*), Speech Synthesizer (*SS*), Face Encoder (*FE*), Residual-

* Corresponding author

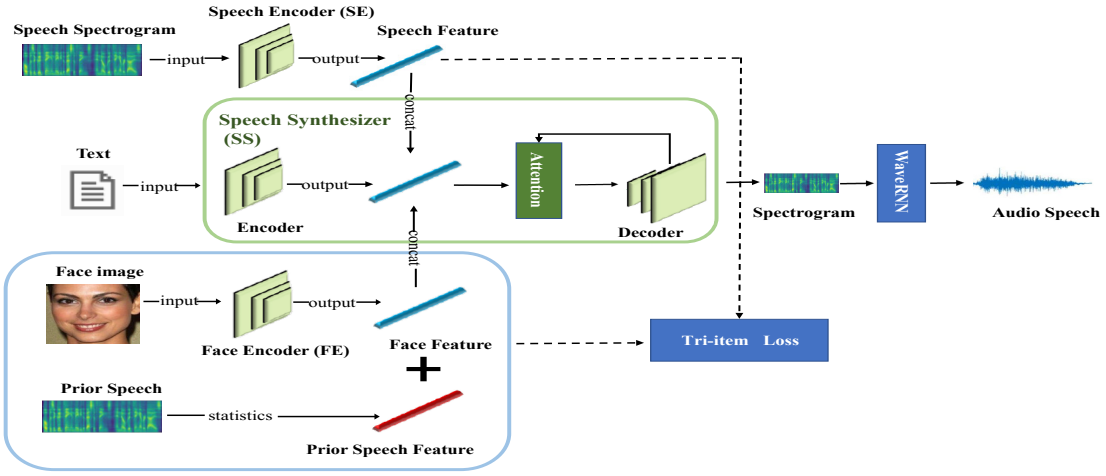


**Fig. 1**. Overview of our FR-PSS with the prior speech.

guided strategy and tri-item loss function, respectively.

### 2.1. Speech Encoder and Speech Synthesizer

*SE* is to extract speech features. Similar to [11], we apply a network that maps a sequence of Mel spectrogram frames calculated from speech to a fixed-dimensional embedding vector. The 40-channel Mel spectrogram is fed to the network, which is composed of three LSTM layers with a total of 768 units, and after each layer, there is a 256-dimensional projection. In the inference process, each utterance of speakers is divided into 800ms windows with 50% overlapping. The network runs independently on each window, and generates the final speech embedding by the average and normalization. We use generalized end-to-end speaker verification losses to train the network.

As for the *SS*, we use Tacotron 2 model, which includes an encoder and a decoder that introduces an location sensetive attention. The details can be referred to [2]. The pretrained WaveRNN in [12] is exploited as the vocoder in this work.

### 2.2. Face Encoder

We use *FE* to extract face features which is used to synthesize speech. It is a four convolution blocks structure, which stacks the convolution kernel of size 1×1, 3×3 and pooling operations. The dimension of the output feature maps after these three operations are the same, so that we can add them in the channel dimension. This can increase the network's width and improve its adaptability to scale. ReLU operation is performed after each convolution layer to increase the nonlinear character of the network. Finally, a 256-dimensional embedding is achieved by using a 1×1 convolution.

To make *FE* owes an ability to focus on face features and ignore the background noise, we utilize a lightweight general purpose module CBAM [13], which can be easily integrated into an end-to-end CNN architecture. CBAM contains a channel and a spatial attention, which are embedded into decoder. Besides, we also tried channel-focused attention mechanism [14], and it turns out that CBAM performs better.

### 2.3. Residual-guided Strategy

Due to natural mismatch between face image and speech, it is hard to extract complete speech features from face image. Therefore, we add a prior information to help *FE* extract speech features. By introducing the prior speech feature, we exploit the idea of the residual to remove the main similar part of the speech (*i.e.*, prior speech feature), thereby highlighting subtle changes depicted by speech feature. It is to reduce the training difficulties and learn more representative speech features. Our FR-PSS converts the face image to speech by a network $\phi$: $\phi(f,t) = SS\left(FE(f) + t + s_{prior}\right)$, where $SS$ is the speech synthesizer, $t$ is the text, $f$ means the image of input face, and $s_{prior}$ is the prior speech feature calculated before the training stage.

Two speech priors are investigated in this work. The first one is neutral speech prior, which is the arithmetic mean of a large gender-balance speech dataset[1]: $s_{prior} = \frac{1}{n}\sum_{i=1}^{n} SE(s)$, where $s$ denotes the audio clips and $n$ is the number of speaker. *SE* is a CNN structure to extract speech feature by taking $n$ = 10, 50, 100, 500, and 1000. In this work, we finally take $n$ equals 500.

The second prior speech feature is gender-dependent prior by assigning two prior speech features to males and females, respectively. To achieve this, a robust classifier network is first needed to predict the gender based on the face image. We use a gender prediction network [15] to predict the gender of speakers. It is trained on the Adience dataset [16] and tested on VGGFace2 dataset [17].

### 2.4. Tri-item Loss Function

In order to make *FE* extract the personalized facial features better, we formulate a tri-item loss function, which is com-

[1]The dataset contains the same number of male and female speakers.

posed of $L_2$ loss, the negative cosine similarity loss and the triplet loss [18].

We consider that $L_2$ loss can be used to measure the error between facial features and speech features and improve the perceptual similarity at the abstract level. The negative cosine similarity is introduced to make speech features and face features vectors similar in direction. The triplet loss is to make the distance between speech feature and the corresponding face feature vectors closer, and make the distance between the speech feature vector and the irrelevant face feature vectors farther. The loss function is given as:

$$L_{Total} = 1 - \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} + \frac{\sum_{i=1}^{n} (A_i - B_i)^2}{n} + max(d(A,B) - d(A,C), 0), \tag{1}$$

where $d$ denotes distance of embedding vectors, $A$ and $B$ are the speech feature and the corresponding face feature embedding vectors. $C$ is the face feature vector of the irrelevant speaker, and $n$ represents the dimension of the feature vector.

## 3. EXPERIMENT

### 3.1. Datasets

Three public datasets are used to train the proposed FR-PSS:

For the *SE*, we use the Voxceleb2 [19], which contains 6000 celebrities that contribute more than one million utterances. We reduce the audio sampling rate to 16 kHz and divide it into three subsets (*i.e.*, training, validation and test). Note that there is no speaker overlapped among these three subsets for all experimental setup in this work.

For the *SS*, we use LibriTTS [20] consisting of two training sets, which contains 436 hours of audio speech uttered by 1172 English speakers, with a sampling frequency of 16 kHz.

As for the *FE*, we use the voice recording from the Voxceleb2 dataset and the face images from the manually filtered version of VGGFace2 [17] dataset. We select an intersection of the two datasets with the common identities and eliminate non-English speakers, deriving 149,354 voice recordings and 139,572 face images of 1089 subjects. The ratio of our training set to our test set is 8:2.

### 3.2. Implementation Details

We introduce the implementation details for *SE*, *SS* and *FE*, respectively.

First, we use the Voxceleb2 dataset to train the *SE*. Our model is implemented by PyTorch 1.1.0 with GPU Tesla K80, and is optimized by Adam with the learning rate of 0.001 and the exponentially decay rate of 0.9 at every 5 epochs. We train our model with 1.56M steps and batch size 64. The number of utterance is 10 per speaker.

Secondly, based on the LibriTTS [20], we first train the speech synthesizer network, and then train the WaveRNN [12] based on the output of the speech synthesizer network. At the speech synthesizer decoder side, the ground truth is passed in rather than the predicted results. The batch size is 64, and we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-6}$. The initial value of the learning rate is $10^{-3}$, and the exponential drops to $10^{-5}$ after 50,000 iterations. The pretrained WaveRNN in [12] is exploited as the vocoder in this work.

Lastly, we use the spectrogram and face-image pairs to train the *FE*. The *FE* is optimized by Adam and the learning rate of 0.01 with the exponentially decay rate of 0.9 at every 5 epochs. We finally train our model with 50 epochs, and the batch size is 8.

### 3.3. Evaluation Metrics

To evaluate the quality of speech synthesized by the FR-PSS quantitatively, we compare speech features extracted from the generated audio and the original audio, using the $L_1$, $L_2$ and Cosine Similarity loss ($Cos$). When $L_1$ and $L_2$ tends to 0 or $Cos$ tends to 1, the difference between two features are similar in the direction. Besides, we rely on Mean Opinion Score (MOS) evaluations based on subjective listening tests. It is to evaluate synthesized speech in two aspects: its naturalness and similarity to real speech from the target speaker.

## 4. RESULT AND ANALYSIS

### 4.1. Ablation Experiment on Loss Function

To verify the effectiveness of the tri-item loss function we proposed, we compare it with different loss functions. The decline of loss functions are shown in Fig. 2. We can see that the convergence speed is obviously accelerated when the proposed tri-item loss function is used to train the model. It shows that the proposed tri-item loss function is effective.

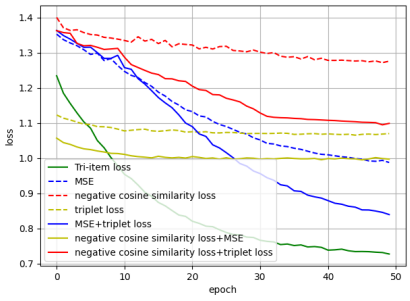


**Fig. 2**. Convergence rate of loss functions.

### 4.2. Quantitative Result

Recall that in the previous work [1, 2, 4, 5], a set of audio data are used to train a model to obtain the personalized speech feature. We call this method that synthesize personalized speech using speaker audio embeddings SYNTH-AUDIO in this work. While the embedding vector of SYNTH-AUDIO is made by applying the *SE* to the audio of the speaker, that of FR-PSS is generated by applying the *FE* to the face images

of the speaker. Since the *FE* was trained to minimize the loss between the embedding vector from the *SE* and output vector from the *FE*, SYNTH-AUDIO can be considered as the upper limit of this framework.

In Table 1, we calculate the differences between speech features extracted from generated audio using different methods and the true speech feature ($\mathrm{T_{speech}}$). We compare their distance by the $L_1$, $L_2$ and $Cos$. It can be seen that the speech obtained by FR-PSS is close to the SYNTH-AUDIO. In particular, adding gender prior knowledge can improve performance significantly. For male priors, we only use male speakers for training and testing, and for female priors we only use female speakers for training and testing. Compared with neutral prior experiments, gender prior avoids gender differences, so it gets smaller errors.

Besides, we compare the face features obtained by the *FE* and $\mathrm{T_{speech}}$ (see Table 2). Results show that by adding prior knowledge, the values of $L_1$ and $L_2$ decrease dramatically, and the $Cos$ approaches to 1.

### 4.3. Qualitative Analysis

A visualization of speakers embedding is shown in Fig. 3 using the U-MAP [21]. We randomly select 10 speakers with 5 utterances per speaker. It can be seen that speakers features are well separated in the speaker embedding space for different speakers, either using the face feature or using the synthesized speech feature by our FR-PSS. Besides, different utterances for the same speaker are well clustered. This shows that the speech synthesized by our model can well represent the personalized speech information.

**Table 1**. Performance of the proposed FR-PSS. $\mathrm{FR-PSS}$, $\mathrm{FR-PSS_n}$, $\mathrm{FR-PSS_f}$ and $\mathrm{FR-PSS_m}$ means FR-PSS with no speech prior, neutral prior, female prior and male prior, respectively. The difference between them and $\mathrm{T_{speech}}$ are shown. $\downarrow$ represents the smaller the better, and $\uparrow$ represents the larger the better.

| **Method** | $L_1 \downarrow$ | $L_2 \downarrow$ | $Cos \uparrow$ |
|---|---|---|---|
| SYNTH-AUDIO | 6.251 | 0.664 | 0.887 |
| $\mathrm{FR-PSS}$ | 8.388 | 0.865 | 0.869 |
| $\mathrm{FR-PSS_n}$ | 7.769 | 0.805 | 0.836 |
| $\mathrm{FR-PSS_f}$ | 6.963 | 0.785 | 0.881 |
| $\mathrm{FR-PSS_m}$ | 7.129 | 0.809 | 0.892 |

**Table 2**. Difference between the face features generated by the *FE* and the true speech features.

| **Method** | $L_1 \downarrow$ | $L_2 \downarrow$ | $Cos \uparrow$ |
|---|---|---|---|
| $\mathrm{FR-PSS}$ | 9.579 | 0.896 | 0.847 |
| $\mathrm{FR-PSS_n}$ | 5.920 | 0.567 | 0.910 |
| $\mathrm{FR-PSS_f}$ | 5.462 | 0.509 | 0.921 |
| $\mathrm{FR-PSS_m}$ | 5.613 | 0.516 | 0.943 |

### 4.4. Subjective Evaluation

MOS is an assessment experiment based on subjective listening test, and it includes two dimensions: naturalness and similarity. To verify the quality of the speech synthesized by our proposed model, we invite 12 listeners to participate in this evaluation. In the naturalness evaluation, each listener rates the naturalness of 100 speech samples. In the similarity evaluation, each listener is given 20 pairs of samples and is asked to score how similar the synthesized speech is to the real one. All our MOS evaluations are aligned to the absolute category rating scale [22], with rating scores from 1 to 5, and the experimental results are shown in Table 3.

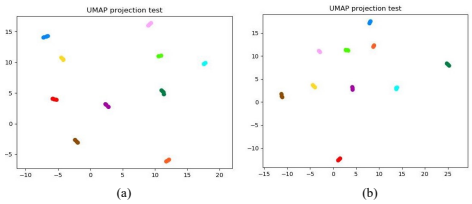


**Fig. 3**. Visualization of speaker embeddings. (a): features of face image. (b): features of synthesized speech by FR-PSS. The same color represents the same person.

**Table 3**. MOS on subjective listening test.

| **Method** | **Naturalness↑** | **Similarity↑** |
|---|---|---|
| SYNTH-AUDIO | 3.83 | 2.43 |
| FR-PSS(proposed) | 3.60 | 2.12 |
| Face2Speech | 3.52 | 1.98 |

Since our work uses face information to synthesize personalized speech, the MOS score is natural slightly lower than SYNTH-AUDIO which is the upper limit of this framework. However, we extract speech features directly from face images, and a good performance is achieved compared with the model that extract speech features from speech to synthesize personalized audio. And in terms of naturalness and similarity, our proposed method is better than Face2Speech. Samples are available at https://github.com/hxs123hxs/FR-PSS_example.

## 5. CONCLUSION

In this work, we innovatively propose the FR-PSS model to extract personalized speech features from human faces and synthesize personalized speech using neural vocoder. By designing three speech priors, a residual-guided strategy is introduced to guide the face feature to approach the true speech feature as much as possible. Experimental results show that our FR-PSS achieves a satisfying performance compared with the method that synthesizes personalized speech using a speech-derived embedding vector. In the future, we will investigate a multi-task framework to further improve the efficiency of the personalized speech synthesis.

## 6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China(No.61977049), National Natural Science Foundation of China(No.62101351) and the Tianjin Key Laboratory of Advanced Networking.

## 7. REFERENCES


[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *in proc. ICASSP*, pp. 4779–4783, 2018.

[3] Li Liu, Thomas Hueber, Gang Feng, and Denis Beautemps, "Visual recognition of continuous cued speech using a tandem cnn-hmm approach.," pp. 2643–2647, 2018.

[4] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," pp. 4485–4495, 2018.

[5] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *in proc. ICLR*, pp. 214–217, 2018.

[6] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, and Wojciech Matusik, "Speech2face: Learning the face behind a voice," *in proc. CVPR*, pp. 7539–7548, 2019.

[7] Jianrong Wang, Xiaosheng Hu, Li Liu, Wei Liu, Mei Yu, and Tianyi Xu, "Attention-based residual speech portrait model for speech to face generation," *arXiv preprint arXiv:2007.04536*, 2020.

[8] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu, "Audio-driven talking face video generation with learning-based personalized head pose," *arXiv preprint arXiv:2002.10137*, 2020.

[9] Shunsuke Goto, Kotaro Onishi, Yuki Saito, Kentaro Tachibana, and Koichiro Mori, "Face2speech: Towards multi-speaker text-to-speech synthesis using an embedding vector predicted from a face image," *in proc. Interspeech*, pp. 1321–1325, 2020.

[10] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[11] W. Li, W. Quan, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," *in proc. ICASSP*, pp. 4879–4883, 2018.

[12] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," *in proc. ICML*, vol. 80, pp. 2410–2419, 2018.

[13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," *in proc. ECCV*, pp. 3–19, 2018.

[14] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *in proc. PAMI*, pp. 7132–7141, 2018.

[15] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," *in proc. CVPRW*, pp. 34–42, 2015.

[16] E Eidinger, R Enbar, and T Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[17] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 2018, pp. 67–74.

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," *in proc. CVPR*, pp. 815–823, 2015.

[19] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018.

[20] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *Proc. Interspeech 2019*, pp. 1526–1530, 2019.

[21] L. Mcinnes and J. Healy, "Umap: Uniform manifold approximation and projection for dimension reduction," *The Journal of Open Source Software*, vol. 3, no. 29, pp. 861, 2018.

[22] ITUT Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, vol. 22, 1996.