

Gender Classification in Voice Signals: An Analysis of Feature Selection and Machine Learning Models

Halavath Peda Sydulu¹, Appalaneni Lavanya², Ippi Sumalatha³

^{1,2,3}Assistant Professor, Department of CSE, Malla Reddy Engineering College and Management Sciences, Hyderabad, Telangana.

Abstract

Human voice carries an abundance of paralinguistic information, making it an invaluable resource for various voice recognition applications. Gender classification is a crucial aspect of voice signal analysis, albeit a challenging task. To distinguish between male and female voices, a range of techniques is applied to identify pertinent features for model development using training data. These models play a key role in determining the gender of a given voice signal. This research offers a comprehensive analysis of well-established voice signal features using a prominent dataset. It explores various machine learning models from different theoretical families for gender classification and employs three feature selection algorithms to identify optimal features that enhance model efficiency. Experimental results emphasize the significance of specific subfeatures in boosting classification model performance. Notably, the best recall value achieved was 99.97% without feature selection, 99.7% for two deep learning (DL) and support vector machine (SVM) models, and 100% for SVM techniques with feature selection.

Keywords: Deep Learning, Voice Recognition, Multilayer Perceptron Networks, Gender Classification.

1. Introduction

The voice of human speech is an effective communication method consisting of unique semantic linguistic and paralinguistic features such as gender, age, language, accent, and emotional state. The sound waves consisting of human voice are unique among all creatures producing sound since every single wave carries a different frequency. Identifying human gender based on voice has been a challenging task for voice and sound analysts who deploy numerous applications including (i) effective advertising and marketing strategies in customer relationship management (CRM) systems which depend on gender interoperability such as the user interface style as well as preferences of words and colours; (ii) investigating criminal voice in crime scenarios; and (iii) enhancing human-computer interaction (HCI) systems especially dialogue systems by customizing services that rely on gender voice and also improving the level of user satisfaction. Because of the importance of identifying gender through voice recognition, the human voice should be converted from the analogue to the digital form to extract useful features and then to construct classification models. The robustness and effectiveness of classifiers are determined by the quality of features that depend on a training set employing machine learning (ML) techniques. Therefore, eliciting voice features plays a vital role in improving the efficiency of classifiers since the human voice is liable for nonuseful features. Research on improving the

efficiency of voice classifiers is copious, particularly studying the process of extracting efficient features from voice including identifying the linguistic content of speech signal components and disposing of nonuseful contents such as background noise. There are a set of features used for recognizing the voice gender. Among the most common features utilized for voice gender recognition are mel-scaled power spectrogram (Mel), mel-frequency cepstral coefficients (MFCCs), power spectrogram chroma (Chroma), spectral contrast (Contrast), and tonal centroid features (Tonnetz). By getting the extracted features combined with the gender label as a form of a training set, ML techniques are used to build a high-quality model for recognizing the voice gender, as shown in Figure 1. In particular, each classification technique is used to build a set of hypothesis models and selects the most optimal one. This model classifies the unknown voice label by receiving the voice features and categorizing the voice gender.

2. Literature Survey:

A multitude corpus of research has been conducted to address the efficiency of voice classifiers aiming to enhance the accuracy of programs being used. Hu et al. [1] used two level classifiers (pitch frequency and GMM classifier) to recognize the speaker's gender on the TIDIGITS dataset achieving a success rate of 98.65%. Djemili et al. [2] used four classifiers including GMM, multilayer perceptron (MLP), vector quantization (VQ), and learning vector quantization (LVQ) to analyze voices taken from IViE corpus. They managed to achieve a 96.4% success rate. Li et al. [3] combined the estimated voice acoustic level of five different methods into one score level. The results were obtained on using the aGender dataset for the gender category with a 81.7% success rate. Yucesoy and Nabiyev [4] proposed a system for identifying speakers using a fusion score of seven subsystems where the feature vectors are the MFCC, PLP, and prosodic on three different classifiers that are GMM, SVM, and GMM-SV-based SVM combined at the score level. The classification success rate on gender identification using the aGender database is 90.4%. Lee and Kwak [5] used two classifiers: SVM and decision tree (DT) with the MFCC feature, on a private corpus identifying gender voice. The overall accuracies using MFCC-SVM and MFCC-DT for gender classification were 93.16% and 91.45%, respectively. The most efficient classifiers and feature extractors of superior accuracy on voice gender recognition include deep neural networks (DNNs) and convolutional neural networks. Qawaqneh et al. [6] proposed an adequate technique to enhance the MFCC features and then adjust the weights between DNN layers. These improved MFCC features are evaluated on DNN and I-Vector classifiers where the overall accuracies are 58.98% and 56.13%, respectively. Sharan and Moir [7] compared two classification techniques (DNN and SVM) using single and combined feature vectors for robust sound classifications. The results showed a better performance for the DNN technique as it is robust and has low sensing to the noise approach.

3. Deep Gender Recognition (DGR)

The proposed methodology for speech gender classification includes a set of stages as briefly discussed. The stages start by converting the voice, from its abstract representation, into a consistent form in order to extract the relevant features. Then, the relevant features are selected

as inputs for building a classifier model for recognizing the gender of a human voice. In addition, a DL model is being built to automatically extract useful features and feed them into a fully connected artificial neural network (ANN) for classification. However, here, a set of process for extracting features for other models rather than DL and classification techniques are summarized as follows.

Voice Preprocessing A transmitted voice is inevitably vulnerable to noise interference and voice attenuation that needs a preprocessing process to purify it for feature extraction. This phase shows a set of steps as follows.

A/D Signal Conversion A/D signal conversion is used to convert the given voice from the analogue to the digital signal by common sampling and quantization techniques [8]. The A/D conversion formulates the signal in an understandable form by machine for easy manipulation.

Preemphasis Process Because of attenuation at high-frequency segments of the voice signal, there is a necessary need to use a preemphasis filter. The preemphasis filter flattens the signal (or speech) waveforms. The process filters low-frequency interference, especially power frequency interference at low-frequency segments, and emphasizes the high-frequency portions in order to produce a high-pass filter to carry out spectral analysis interference. This process occurs after A/D conversion by the first-order digital preemphasis filter.

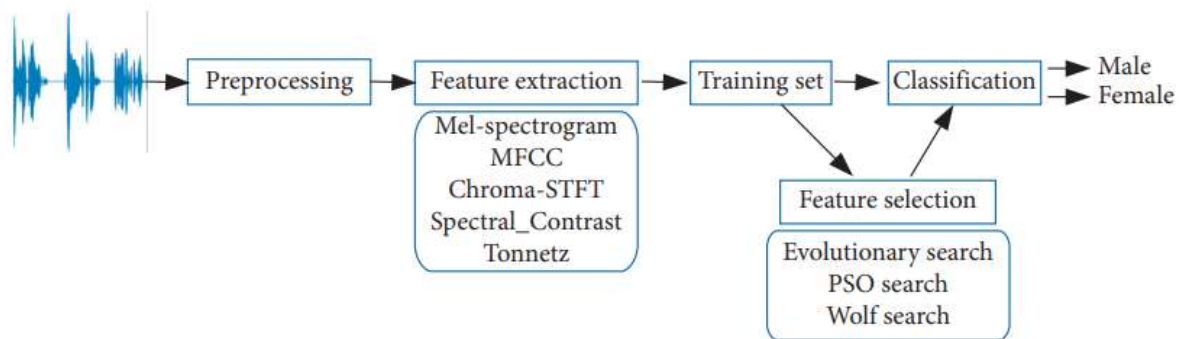


Figure 1: General gender recognition framework.

MFCC MFCC represents accurately the vocal tract that is a filtered shape of a human voice and also manifests itself in the envelope of a short-time power spectrum, as shown in Figure 2(c). In order to compute MFCCs, a set of sequential steps should be followed:

- (1) Framing the Signal into Short Frames. The audio signal is framed into 20–40 ms (25 ms is standard) frames to overcome changes in the sample in a short time period as it is constantly changed in a long period of time.
- (2) Periodogram of Power Spectrum. This calculates for each frame the periodogram estimation of the power spectrum, which identifies the frequencies in the frame.
- (3) Applying the Mel Filterbank to the Power Spectra (or Summing the Energy in Each Filter). A filter is required for estimating the energies in various frequency regions that

appear in a group of aggregated periodogram bins because of unnecessary information in periodogram spectral estimation. Hence, the mel filterbank estimates the energy near 0 Hz and then for higher frequencies as there is less concern for variations.

(4) Logarithm of All Filterbank Energies. Large variations of energies are scaled using a logarithmic scale as there are no different sounds in large energies. The logarithmic scale is a channel normalization technique that is also exploited for cepstral mean subtraction.

(5) DCT of the Log Filterbank Energies. Because of the correlation in filterbank energies that lead to overlapping, the DCT is used to decorrelate the energies. This generates diagonal covariance matrices as features.

(6) 2-13 DCT Coefficients. Higher DCT coefficients are chosen to reduce the fast changes in the filterbank energies and discard the rest.

Chorma-STFT (Short-Time Fourier Transform) Chorma-STFT computes a chromagram from a waveform or power spectrogram, Chroma features are powerful representatives for a music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or Chroma) of the musical octave.

Spectral_Contrast Spectral_Contrast computes spectral contrast, using the method defined in [11]. It represents the relative spectral distribution instead of the average spectral envelope.

zTonnetz computes the tonal centroid features (or Tonnetz), following the method in [1] that detects changes in the harmonic content of musical audio signals.

Classification Learning Techniques Classification learning algorithms aim to find an optimal classifier model for recognizing test samples of provided features and unknown labels. Several learning techniques fundamentally reveal a philosophical theory in modelling knowledge as a mathematical form. In order to cover the diversity in using different forms, a set of classification learning algorithms of various families are used. In particular, the selected classifiers in terms of the family include the following [12].

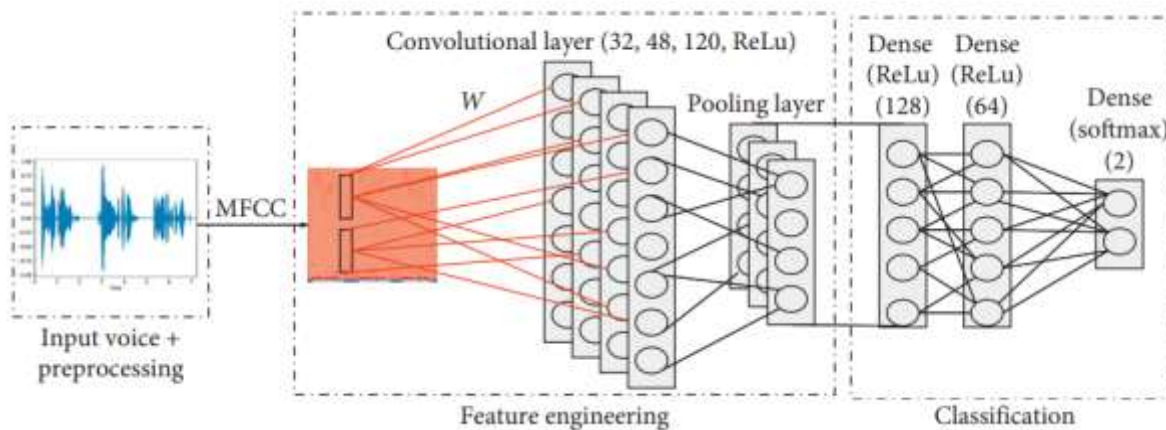


Figure 2: One-dimensional conventional neural network.

DNN is a framework of two phases, which are feature engineering and classification [13, 14]. The feature engineering process automatically extracts useful and nonlinear features from the raw data using convolutional and pooling layers by optimizing the weights W (or feature maps) between layers [15]. In the classification phase, the useful features are flattened as a vector to be fed into a fully connected ANN. In this work, the architecture of the DNN, as shown in Figure 3, receives the MFCC features of the input voice as one-dimensional (1D) data. These features are then fed into a convolutional layer that consists of three layers of 32, 48, and 120 neurons using ReLu as a nonlinear activation function. A pooling (or subsampling) layer follows the conventional layers using max function to reduce the size of resulted features. Finally, such features are flattened as the input vector to a fully connected ANN which is three dense layers of 128 neurons, 64 neurons using ReLu function, and 2 output neurons representing the gender of the input voice using softmax function (i.e., a normalized probability function). Two 1D DNN models are used which are the normalized deep convolutional neural network (DL_norm) and deep convolutional neural network (DL). The parameter settings of the DNN are 1000 epochs (or the number of iterations), 25% dropout (or regularization), Adam optimizer, and pooling and feature map size of 2×2 . Lazy learners simply classify a new sample by estimating the vector similarity between the sample features and the vectors of samples in the training set and then assign the label of the most similar ones to that test sample. Lazy classifiers differ from other methods known as eager learners. Eager learners construct a machine learning model before the testing process as ready-to-use classifier models. The lazy learners used in this study are IBk and KStar (K*).

Bayes is a direct approach that finds the best hypothesis by using Bayes' theorem as a probability theorem for building rule- or graph-based classification models. Two well-known methods are used, which are the Bayesian network (BN) and naive Bayes (NB) models.

Functions. In this family, the classifier builds a function (or hypothesis) of the input domain (i.e., features) and maps it into a range of outputs (i.e., labels) to form a function for classification. A set of models are used which are multilayer perceptron (MLP), SMO (sequential minimal optimization for SVM), logistic (L), support vector machine (SVM linear (S_L), SVM polynomial (S_P), and SVM radial (S_R)), and latent Dirichlet allocation (LDA).

Feature Selection Techniques Building an optimal classifier model is affected by no-relevant features used for constructing such a model. These features drive the model to produce low accuracy for provided labels that leads to the underfitting or overfitting problem. Therefore, the necessity for selecting the relevant subset is needed. Three feature selection optimizers are used which are derived from the natural behaviour—evolutionary search, particle swarm optimization (PSO) search, and wolf search [16–18]. Each algorithm generates a set of individual solutions and then selects the optimal solution based on an evaluation metric and a learner optimizer (or evaluator). In this work, the evaluation metric used is area under the ROC curve (AUC) to validate whether a classifier can separate positive and negative samples and identify the best

threshold for separation [19]. On the contrary, the RF classifier from the tree family is used as an evaluator to select the best subset features.

Evaluation In this phase, particularly in the training phase, the 10-fold cross-validation method is used for each experiment by repeating it 10 times at each process of building a classifier. The evaluation metrics used are precision and recall [20]. Precision is the ratio of relevant samples to the retrieved ones, while recall is the ratio of retrieved and relevant samples to the total amount of relevant samples

4. RESULTS

Implementation of Gender identification uses the following technologies. PyAudio provides Python bindings for PortAudio, the cross-platform audio I/O library. With PyAudio, you can easily use Python to play and record audio on a variety of platforms. PyAudio is inspired by: ... tkSnack. A Fast Fourier transform (FFT) is an algorithm that computes the discrete Fourier transform of a sequence, or its inverse . Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa



Figure 3: input gui for configure



Figure 4: recording the voice window

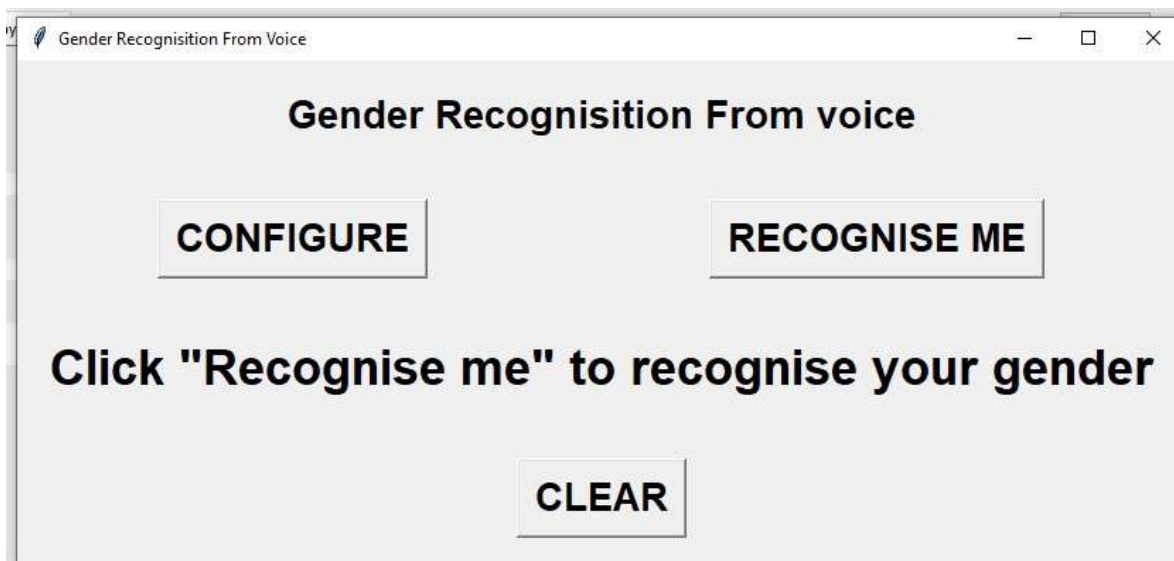


Figure 5: recognize of the voice window

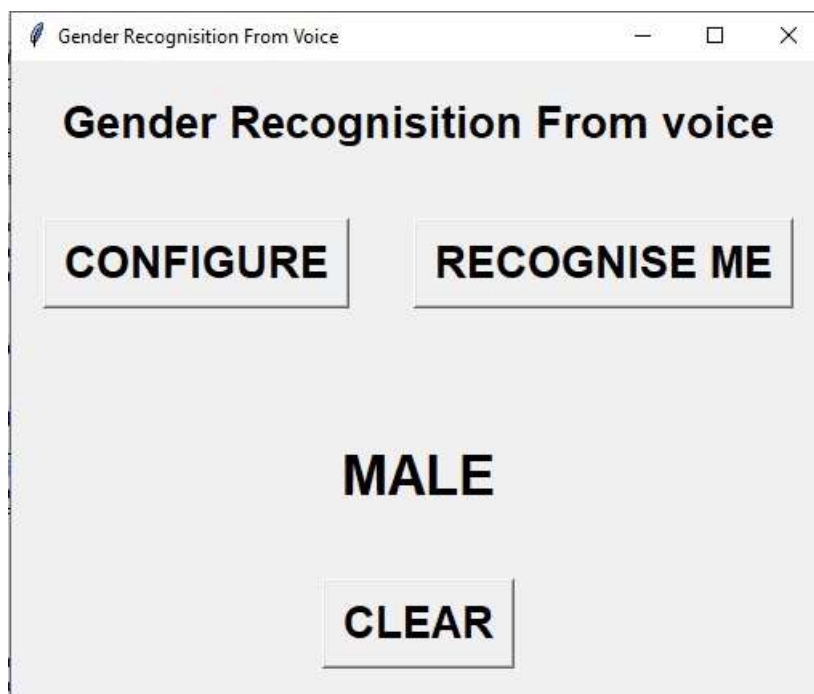


Figure 6: output window

Conclusion and Future Work

Recognizing the gender of human voice has been considered one of the challenging tasks because of its importance in various applications. The contributions are threefold including (i) studying the extracted features by examining the correlation between each other, (ii) building classification models using different ML techniques from distinct families, and (iii) evaluating the natural feature selection techniques in finding the optimal subset of relevant features on classification performance. In particular, three feature categories perform efficiently because of

their theoretical methodology in extracting the relevant coefficient energies in the voice signal, which are the MFCCs, Chroma, and Mel, and this answers the research question RQ1. From the perspective of the performance of classifiers, the ML techniques behave in different ways. The results showed that the function family gained better performance compared to the other families. Although the function family has superior results, the other techniques have promising results, and this leads us to answer the research question RQ2. Finally, a set of experiments are conducted using three common feature selection techniques inspired by nature, which are EA, PSO, and wolf methods using the RF as an evaluator. These wrapper selection techniques select subfeatures from the feature categories which are on average approximately 39.1% on overall features. In spite of a small number of subfeatures, the performance of ML techniques was increased as there are some features that are not relevant in determining the gender of human voices. This also answers the research question RQ3. In the future work, more experiments are being conducted to use many feature categories, ML techniques, and other natural feature selection techniques. Furthermore, the proposed techniques are being examined on different datasets since here only a standard artificial voice from the study in [21] is used. The reasons behind using it are that it contains many different languages (i.e., 20 languages), as well as the voice text is too long.

References

- [1] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," *Security and Communication Networks*, vol. 5, no. 2, pp. 211–225, 2012.
- [2] R. Djemili, H. Bourouba, and M. C. A. Korba, "A speech signal based gender identification system using four classifiers," in *Proceedings of the 2012 International Conference on Multimedia Computing and Systems*, pp. 184–187, Tangiers, Morocco, May 2012.
- [3] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.
- [4] E. Yücesoy and V. V. Nabiyev, "A new approach with scorelevel fusion for the classification of a speaker age and gender," *Computers & Electrical Engineering*, vol. 53, pp. 29–39, 2016.
- [5] M.-W. Lee and K.-C. Kwak, "Performance comparison of gender and age group recognition for human-robot interaction," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 3, no. 12, 2012.
- [6] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," *Knowledge-Based Systems*, vol. 115, pp. 5–14, 2017.
- [7] R. V. Sharan and T. J. Moir, "Robust acoustic event classification using deep neural networks," *Information Sciences*, vol. 396, pp. 24–32, 2017.

- [8] J. G. Proakis and D. G. Manolakis, "Digital signal processing," in Principles, Algorithms, and Applications, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.
- [9] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 6, pp. 1097–1111, 2008.
- [10] K.-I. Kanatani, "Fast fourier transform," in Particle Characterization in Technology, pp. 31–50, CRC Press, Boca Raton, FL, USA, 2018.
- [11] W. Abdulla, N. Kasabov, and D.-N. Zealand, "Improving speech recognition performance through gender separation," Changes, vol. 9, p. 10, 2001.
- [12] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.
- [13] W. Di, A. Bhardwaj, and J. Wei, Deep Learning Essentials: Your Hands-On Guide to the Fundamentals of Deep Learning and Neural Network Modeling, Packt Publishing, Birmingham, UK, 2018.
- [14] Z. Li and D. Hoiem, "Learning without forgetting," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 12, pp. 2935–2947, 2018.
- [15] S. Hinterstoisser, V. Lepetit, P. Wohlhart, and K. Konolige, "On pre-trained image features and synthetic images for deep learning," in Proceedings of the European Conference on Computer Vision, pp. 682–697, Springer, Munich, Germany, September 2018.
- [16] Y. Chtioui, D. Bertrand, and D. Barba, "Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision," Journal of the Science of Food and Agriculture, vol. 76, no. 1, pp. 77–86, 1998.
- [17] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, "Feature subset selection approach by gray-wolf optimization," in Afro-European Conference for Industrial Advancement, A. Abraham, P. Krömer, and V. Snasel, Eds., pp. 1–13, Springer International Publishing, Cham, Switzerland, 2015.
- [18] B. Xue, M. Zhang, and W. N. Browne, "Particle Swarm Optimization for feature selection in classification: a multiObjective approach," IEEE Transactions on Cybernetics, vol. 43, no. 6, pp. 1656–1671, 2013.
- [19] L. B. Lusted, "Signal detectability and medical decisionmaking," Science, vol. 171, no. 3977, pp. 1217–1219, 1971. Scientific Programming 11
- [20] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in Proceedings of the 23rd International Conference on Machine Learning (ICML '06), pp. 233–240, ACM, New York, NY, USA, 2006.

[21] ITU-T Recommendation P.50, "Objective measuring apparatus," in Proceedings of the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T) , Geneva, Switzerland, September 1999.