Text-dependent Speaker Recognition by Combination of LBG VQ and DTW for Persian language

Mahdi Keshavarz Bahaghighat
Electrical and ComputerDepartment,
Raja University,
Qazvin, Iran.

Farshid Sahba
Electrical and Computer Department,
Raja University,
Qazvin, Iran.

Ehsan Tehrani
Electrical and Computer Department,
Islamic Azad University
Buinzahra branch, Buinzahra, Iran.

ABSTRACT
This paper gives a novel approach of automatic speaker recognition technology, with an emphasis on text-dependent speaker recognition. Speaker recognition has been studied actively for several decades. In fact, Speaker recognition system may be viewed as working in four stages, namely, analysis, feature extraction, modeling and testing. After some preprocessing modules, we apply MFCC, as one of the most important feature extraction methods in this field of works, to speech signals independently in order to extract feature vectors. Afterwards, obtained vectors are used by training system to find codewords for ten users in our Persian database by LBG VQ. Finally, we use DTW technique for recognizing a speaker among all. Our experience strongly indicates that the identification rate over 96% can be achieved by the proposed algorithm.

General Terms
Speeches analyze,Speaker recognition.

KEYWORDS
Speaker recognition systems, MFCC, LBG VQ, DTW.

1. INTRODUCTION
Speaker recognition is defined as automatic identification of a speaker based on individual information on speech signal[1,2] .Speaker recognition systems have a large set of applications in everyday life: Time and Attendance Systems, Access Control Systems, Telephone-Banking, Biometric Login to telephone aided shopping systems, Information and Reservation Services, Security control for confidential information and Forensic purposes[5].
Speaker recognition systems are classified as text-dependent (with knowledge of the lexical content) and text-independent (without that knowledge). The text-dependent systems require a user to re-pronounce some specified utterances, usually containing the same text as the training data, like passwords, card numbers, PIN codes, while there is no such constraint in text-independent systems.Therefore, a Common text-dependent (TI) recognition task is vocal password access control in which each user has a private predefined vocal password. Text-independent recognition tasks such as call routing usually do not make any assumptions on the lexical content of the call. It is well known that given thesame experimental conditions (amount of data for training/testing, noise conditions, etc.) text-dependent speaker recognition is more accurate than text-independent recognition.
Speaker models in the TI methods are distributions in the feature space, modeled by VQ codebooks [17,18] built from the extracted features from training speech data. During the test, the TI speaker recognition system tries to find which speaker model (distribution) the test feature-vector-set came from. Such distributions are often overlapping, especially if the password phrases are same or similar for all speakers, leading to lower performances of the TI systems. Text-dependent (TD) speaker recognition methods [15,16]on the other hand exploit the feature dynamics to capture the identity of the speaker. TD methods assume the utterance of a certain password by the speaker and compare the feature vector sequence of the test utterance with the "feature dynamics-model" of all the speakers. Such feature-dynamics based models can be the stored templates of feature vector sequence as used in the TD method using DTW [16] or they can be HMMs trained by a large number of passwords uttered by the speaker [15]. The way a person speaks a certain phrase, captures a lot of his/her speaking style (i.e. the identity), in the co-articulation of various sound units. This important aspect of speaker identity is captured by TD systems and therefore TD systems typically offer much higher performance than the TI systems.

The organization of this paper is as follows:
The speaker recognition system is presented in Section 2 from 2.1 to 2.6. Section 3 describes making reference model while section 4 presents feature matching and final decision algorithm. Finally, conclusions will be remarked in section 5.

2. SPEAKER RECOGNITION SYSTEM
Figure1 illustrates a general speaker recognition system which consists of several main parts including preprocessing, feature extraction, reference model, pattern matching and decision criteria.

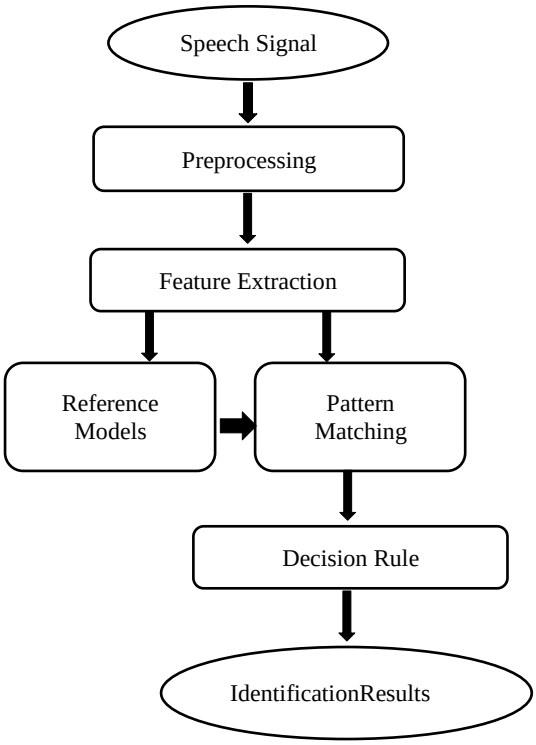


Fig1: General diagram of any speaker identification system

2.1 Preprocessing

Analog speech signal is sampled and quantized using an analog to digital converter to be in digital form. There are some key parameters that should be set optimally such as sampling rate and quantization bits for digital storage.

In order to avoid aliasing problem occurred during sampling and to obtain precise values of digitized samples, usually 8-20 kHz sampling rate and 12-16 bit resolution are considered [8].

After data acquisition process, the next step is to enhance an input speech signal to has higher quality and appropriate characteristic for following processing steps. Preprocessing covers digital filtering, endpoint detection, and time normalization. Filtering is to filter out any surrounding noise using several digital filters.

Endpoint detection is a process of clamping only a desired speech interval. There are a lot of endpoint detection algorithms such as energy-based, zero-crossing, and fundamental frequency have been proposed for speech processing tasks and we used energy-based endpoint detection in this research due to its simplicity. Furthermore, we eliminate 50Hz microphone noise by applying Gaussian low pass filter with cutoff frequency around 200Hz before endpoint detection. In addition to noise reduction and endpoint estimation, time normalization is another step of reprocessing module to stretch or press original speech to normalized speech with desired time duration. This sub-procedure is an optional phase and is just applied to some recognition strategies.

2.2 Feature extraction

The main purpose of this module is conversion of the speech waveform to some type of parametric representation at a considerably lower information rate for further analysis and processing for speaker identification.

The speech signal is a slowly time-varying signal so-called quasi-stationary. When examined over a sufficiently short period of time between 5 and 100ms, its characteristics are fairly stationary. However, over long periods of time for example on the order of 1/5 seconds or more the signal characteristic vary to reflect the different speech sounds being spoken. Therefore, the short-time spectral analysis is the most common way to characterize the speech signal. A wide range of techniques such as Linear Prediction Coding (LPC)[14], Mel Frequency Cepstrum Coefficients (MFCC)[4], and others are used to parametricrepresentation of speech signal for the speaker recognition task. Among all of them, MFCC is perhaps the best known and most popular [4], and it has been used in our work.

MFCC is based on the known variation of the human ear's critical bandwidth frequencies; a bank of filters that spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the Mel-frequency scale, linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz(See Figure5(a)). The structure of MFCC module is shown in Figure2 and we will evaluate it in more details in the flowing stages.

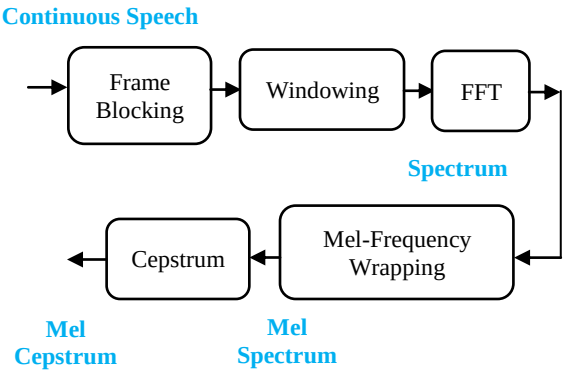


Fig2: The main structure of MFCC processor

2.3 Frame blocking

It can be viewed in Figure3 that input speech signal is blocked into predefined frames with N samples. Each adjacent frame has N-M overlap samples. In fact, after M samples of current frame next frame will begin. So for example, 2nd frame effectively shares 2(N-M) overlap samples with 1st and 3rd frames. Then this process will continue until all the speech is considered within one or more consecutive frames.

According to sampling rate and time slot restriction for stationary speech signal around 30ms and in order to reduce timing process, amounts 256 and 120 are assigned by us to N and M respectively.

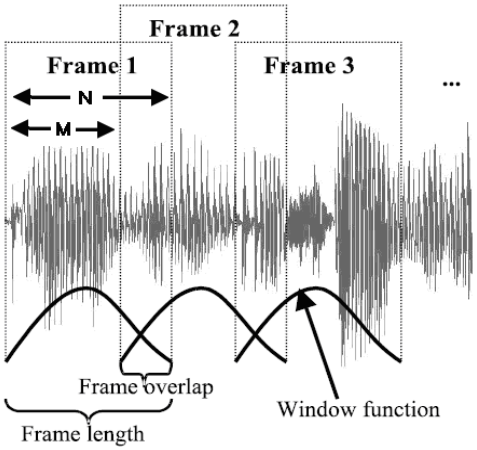


Fig3: Frame blocking and windowing for given speech signal

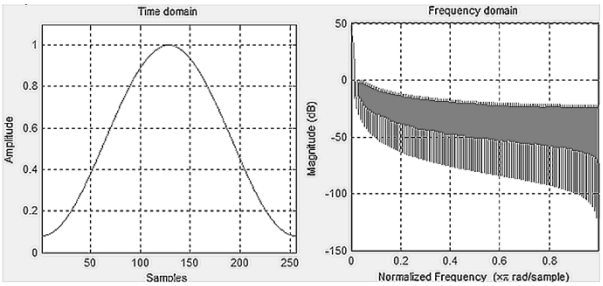
2.4 Windowing

In this stage, all frames should be windowed individually to decrease signal discontinuities at both start and end points of each frame. This is an essential step to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. In our work, Hamming window is used which has the form of Eq1 :

$$W(n) = 0.54\text{-}0.46\cos(2\pi n/(N-1)) \qquad (1)$$

$$0 \le n \le N-1$$

Where N is the number of samples in each frame. Both time and frequency responses of Hamming window are shown in Figure4 for N=256.



(a) (b)

Fig4: Display the 256-points Hamming window and its FFT

So as it has been indicated by Eq2 , the result of windowing x(n) will be the signal y(n) :

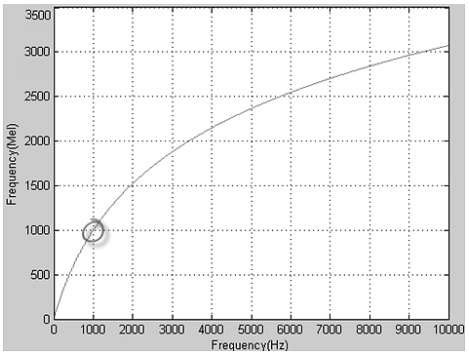
$$y(n) = x(n)W(n), \quad 0 \le n \le N\text{-}1 \qquad (2)$$

2.5 Mel-frequency

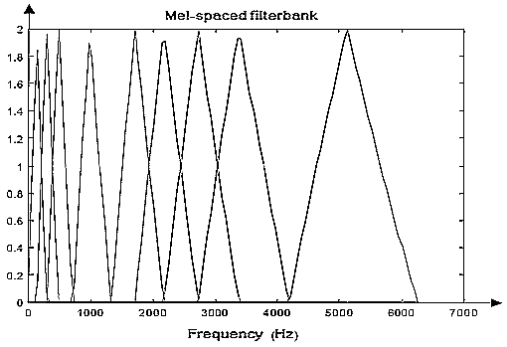
The Mel scale, named by [21] is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and normal frequency measurement is defined by assigning a perceptual pitch of 1000 Mels to a 1000 Hztone(See Figure5(a)), 40 dB above the listener's threshold. Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments. As a result, four octaves on the hertz scale above 500 Hz are judged to comprise about two octaves on the Mel scale. A popular formula to convert frequency in hertz into Mel is given by Eq.:

$$mel(f) = 2595\log(1 + f/700) \qquad (3)$$

After windowing speech signal with the Hamming window, the Fast Fourier Transform should be computed. The magnitude is then weighted by a series of filter frequency responses where center frequencies and bandwidths match those of the auditory critical band filters. These filters follow the Mel-scale whereby band edges and center frequencies of the filters are linear for low frequency (approximately frequencies below 1KHz) and logarithmically increase with increasing frequency as shown in Figure5. These filters are called the Mel-Scale Filter (MSF) bank.Figure5(b) shows the MSF bank with K triangularly shaped frequency responses, which approximate the actual auditory critical band filters that cover the 4 kHz range.



(a)



(b)

Fig5: (a) Mel-scale frequencies vs. frequencies in Hz(b)A sample Mel-scale triangular filter bank

2.6 Calculating Cepstrum coefficients

Now, the log Mel spectrum is converted back by Discrete Cosine Transform (DCT) to time domain. Obtained results are called MFCC. The cepstral representation of the speech spectrum provides a well representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the DCT. Consequently, providing that we define those Mel power spectrum coefficients that were output results of the preceding step as $\tilde{S}_k$,where $k = 1, 2, ..., K$ and K is the number of Mel spectrum coefficient which is defined at 20 , then we will be capable to determine the MFCC according to Eq4 :

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k)\cos[n(k-1/2)\pi/K] \qquad (4)$$

$$n = 1, 2, ..., K$$

In addition, it is worthy to mention that the first coefficient, C_0, is excluded from the DCT because it represents the mean value of input signal which doesn't have noticeable speaker specific information.
So far, theremarked procedure was applied to each speech frame and consequently a set of Mel-frequency Cepstrum coefficients were then computed. This set of coefficients is called feature vectors.
The next section describes how these feature vectors are used by Vector Quantization (VQ) technique to make voice characteristic models of all speakers.

3. REFERENCE MODEL

In order to make reference model for our speaker recognition system, we applied Vector Quantization (VQ) technique that has wide range of applications for voice recognition, due to its easy implementation and high accuracy. It employs the process of clustering between large numbers of given vectors. It can be viewed in Figure6 that VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by the centroid called codeword[9,10]. The collection of all codewords consists of the corresponding codebook.

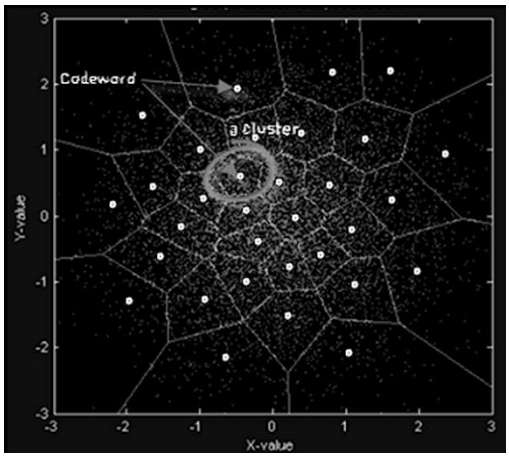


Fig6: An example of a 2DVQ

A model is the set of centroids $\{C_1, C_2, ..., C_M\}$. An entry in this code book is the model of each speaker and let assume that there are M vectors, $\{x_1, x_2, ..., x_M\}$ (in a speaker recognition system there are as many vector as frames in the utterance)and each vector has k components (as same as the number of Mel Frequency Cepstrum Coefficients).
The LBG VQ design algorithm is an iterative algorithm [20] which alternatively solves optimality criteria. The algorithm requires an initial codebook. The initial codebook is obtained by the splitting method. In this method, an initial codevector is set as the average of the entire training sequence. This codevector is then split into two vectors. The iterative algorithm is run with these two vectors as the initial codebook. The final two codevectors are split into four and the process is repeated until the desired number of codevectors is obtained. The algorithm is summarized in the flowchart of Figure7.

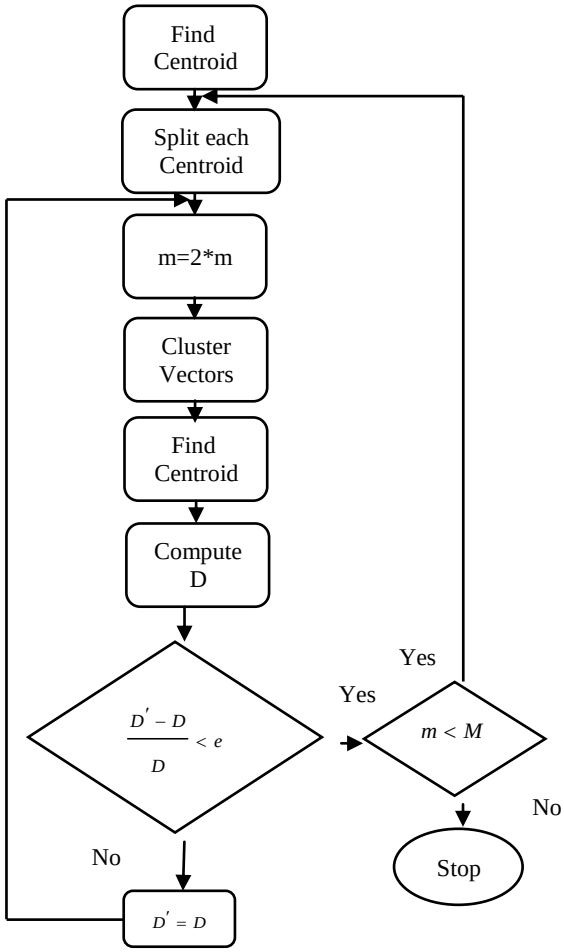
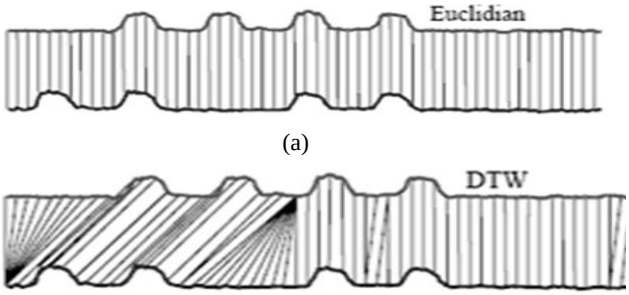


Fig7: Flowchart of LBG-VQ algorithm

4. FEATURE MATCHING AND FINAL DECIESION

In the preceding section, the feature vectors of each speaker were calculated separately and individually. During training system to find centroid of each cluster we face the problem of non-equal speech wave time lengths. It can be solved by zero padding to training sequence of each speaker to reach same size, but after estimation the codewords of all speakers independently, we may have L non-equal code lengths as a result of which simple metric techniques such as Euclidean distance will not be able to distinguish between the unknown word and codebooks accurately. To cope with this major problem, it is necessary to apply adaptive feature matching approaches [10, 13] with appropriate correlation index.
Dynamic Time Warping (DTW) [3, 7] is a much more robust distance measurement method for time series, allowing similar shapes to match even if they are out of phase in the time axis. This fact clearly is shown in Figure 8. Therefore, we use DTW for our purpose to strongly increase the identification rate of our system. We calculate the distance between feature vector which is extracted from current unknown speaker and all of our produced reference models by DTW algorithm and finally we will make our decision about valid speaker easily by minimum distance criteria.



(a)

(b)

Fig8: Draw a comparison between
(a)Euclidean Distance:
Sequences are aligned "one to one".
(b) Warped Time Axis:
Nonlinear alignments are possible.

5. CONCLUSION

In this literature, we have proposed novel approach of text-dependent speaker recognition system for Persian language based on combination of two robust algorithms, VQ and DTW. Both of them have their pros and cons, so the principal reason of doing that is to integrate advantages of two algorithms together to not only coping with the various time lengthsproblem in speaker recognition system but also increasing its identification rate.

We have made our database including ten speakers. Each of them had said and repeated the three words Persian phrase "طبیعت زیباست" , which means "the nature is beautiful" , more than 30 times and about 300 speech waves have been collected to train and test the performance of proposed system for ten users. The samples were sampled by 16-bit ADC and two different sample rates, 8KHz and 16KHz. The obtained results indicate achieving over 91% and 96% identification rate for mentioned sample rates, respectively.

6. ACKNOWLEDGEMENT

This study has been supported by Raja University, Qazvin, Iran.

7. REFERENCES

[1] E. Karpov, "Real-Time Speaker Identification " , University of Joensuu, Department of Computer Science, Master's Thesis,2003.

[2] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", ICASSP 2002, pp 4072-4075.

[3] Eamon.Keogh , "Exact indexing of Dynamic Time Warping",2002 Computer Science & Engineering Department Riverside ,university of California, CA92521.

[4] Z. Bin, W. Xihong, C. Huisheng, "On the Importance of Components of theMFCC in Speech and Speaker Recognition", Center for Information Science, Peking University, China, 2001.

[5] Gerrit C. van der Veer, Hans van Vliet," The Human-Computer Interface is the System: A Plea for a Poor Man's HCI Component in Software Engineering", Curricula. CSEE&T 200.

[6] Venayagamoorthy GK, Sunderpersadh N, "Comparison of Text-Dependent Speaker Identification Methods for Short Distance Telephone Lines using Artificial Neural Networks", International Joint Neural Networks Conference (IJCNN 2000), Como, Italy, 24 – 27 July, 2000, vol.5, pp. 253-258 .

[7] Park, S.,Chu, W.,Yoon,J& Hsu, C.,(2000). "Efficient search for similar subsequence of different lengths in sequence database".In Proc.16th IEEE Intconf.on data Engineering.pp.23-32.

[8] C. Wutiwiwatchai,V.Achariyakulporn&C.Tanprasert ,"Text-dependent Speaker Identification using LPC and DTW for Thai Language",1999,Engineering Laboratory, National Electronics and Computer Technology Center,Nation Science and Technology THAILAND.

[9] T. Kinnunen, I. Karkkainen, P. Franti: "Is Speech Data Clustered? - Statistical Analysis of Cepstral Features", Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001),vol. 4, pp. 2627-2630.

[10] T. Kinnunen, T. Kilpelainen, P. Franti: "Comparison of Clustering Algorithms in Speaker Identification", Proc. IASTED Int. Conf. Signal Processing and Communications (SPC 2000), pp. 222-227, Marbella, Spain, September 19-22,2000.

[11] L. Rabiner and B.-H. Juang, "Fundamentals of Speech Recognition",Englewood Cliffs (N.J.), Prentice Hall Signal Processing Series, 1993.

[12] S. Molau, M. Pitz, R. Schluter, H. Ney, "Computing Mel-Frequency CepstralCoefficients on the Power Spectrum", Acoustics, Speech, and Signal Processing,2001 IEEE International Conference, Volume: 1, 2001, pp. 73-76.

[13] L. Liao, M. Gregory, "Algorithms for Speech Classification", ISSPA 1999,Brisbane, Australia.

[14] D. O'Shaughnessy, " Linear Predictive Coding", IEEE Potentials -- Vol. 7,1988, no.1, p.29-3.

[15] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," Proc. ICASSP, pp. II-391-394, (1993).

[16] V. Ram, A. Das, and V. Kumar, "Text-dependent speaker-recognition using one-pass dynamic programming",Proc. ICASSP'06, (2006).

[17] F.K.Soong, A.E. Rosenberg, et al, "A vector quantization approach to speaker recognition", AT&T Tech. Journal, Vol 66, pp 14-26 (1987).

[18] A. Das & P. Ghosh, "Audio-Visual Biometric Recognition by Vector Quantization", IEEE SLT-06, 2006.

[19] Amitava Das &GokulChittaranjan, "Text-Dependent Speaker Recognition by Efficient Capture of Speaker Dynamics in Compressed Time-Frequency Representations of Speech", submitted to Inter-speech 2008.

[20] Y. Linde, A. Buzo& R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.

[21] Stevens, Stanley Smith; Volkman; John; & Newman, Edwin (1937). "A scale for the measurement of the psychological magnitude pitch". Journal of the Acoustical Society of America 8 (3): 185–190.