A Real-Time Speech–Music Discriminator*

RONALD M. AARTS, *AES Fellow,* **AND ROBERT TOONEN DEKKERS**

Philips Research Laboratories, 5656 AA Eindhoven, The Netherlands

A real-time circuit is described which automatically discriminates between speech and music signals. An output of the circuit gives, by means of a fuzzy feature combiner, an estimate of the probability that the input is speech. The discriminator is tested (for both sexes) for various languages, such as English, Danish, Dutch, French, German, and Japanese, against various types of music, such as pop, opera, romantic, baroque, and various solo musical instruments. The discriminator is found to be extremely reliable, the false alarm probability (inferring speech while the input is music) being virtually zero.

0 INTRODUCTION

Automatic classification of audio signals into categories, such as pop, romantic, baroque, and speech, could be an attractive feature for audio and television products. For speech signals the system could switch to a reduced bandwidth to increase the signal-to-noise ratio or to obtain a better speech intelligibility. Another application might be to direct all the speech to a center loudspeaker. If music is classified properly, then it can be equalized to personal preference in combination with the existing room acoustics.

1 PROBLEMS IN AUDIO TRANSMISSION

The problems arising from the reproduction of audio signals are depicted in Fig. 1. The direct speech or music signal d_1 is corrupted by acoustical noise an_1 and an indirect sound id_1 due to a reflective surface rs_1. The sum of these signals S_t is received by a microphone Mic and transmitted (block Tr). Imperfections of the recording and playback system may be modeled by the block Tr as well. Owing to noise in all amplifiers during transmission and reception this signal is corrupted, the corruption being modeled by a lumped noise source n. Besides the direct sound d_2, part of the transmitted sound is reflected yielding id_2, and together with an acoustical noise an_2 these signals are finally perceived by the listener.

1.1 Signal Enhancement

There are various applications for a speech–music enhancer, but its main application is to enhance the speech for television, for example, by automatically switching off the bass boost in the case of a speech signal, or to equalize the signal in the case of music signals. Such a system can be made from two circuits, as shown in Fig. 2. The first part is the discriminator,[1] which determines whether the incoming signal is speech or music, and this will be the main topic of this engineering report.

The second part is the actuator, which can filter or compress the audio signal. However, this will not be discussed here. An important point to consider is how one desires the actuator to function. The aim may be that the signal perceived by the listener resembles d_1 or S_t as closely as possible. On the other hand one can demand for speech optimal speech intelligibility and for music just a "nice" sound. Depending on the position where a speech–music enhancer measures the corrupted signal, the associated noise can be made less annoying. If, for example, the signals id_2 and an_2 must be made less annoying, a microphone at the listener's position must be used. These affairs will strongly determine the final structure of the whole system. The discriminator can change the relevant parameters for the actuator and derives from both the audio signal and the auxiliary input an estimate of the probability given that the input signal is speech. The auxiliary input can be advance information about the repertoire, a source selector, or a signal that is present when one switches to a different TV station. Since the input signal can be either speech or music, there are two possible outcomes of the discrimi-

* Manuscript received 1997 October 30; revised 1999 May 5.

[1] We use the term "discriminator" when there are only two possible decisions, either speech or music, and "classifier" if there are more than two classes, such as speech, pop, opera, romantic, or baroque.

nator, and each decision can be represented by one of the quadrants in the stimulus-response matrix, as shown in Fig. 3. The nomenclature used in this figure is common to the signal detection theory as well as the yes–no tasks in the listening tests, and will be used in the following.

2 CLASSIFICATION PROBLEM

The ultimate goal of the classification of audio signals is a pattern recognizer, which can contain three parts: a transducer, a feature extractor, and a classifier. The transducer converts the input signal to a form suitable for machine processing. The feature extractor (also called receptor, property filter, or attribute detector) extracts presumably relevant data from the input. The classifier uses this information to assign the input data to one of a finite number of categories. The transition between the feature extractor and the classifier is arbitrary. An ideal classifier makes the job of a feature extractor trivial, and vice versa; the distinction is for practical reasons only. A neural net [1], [2] can be very useful as a classifier.

2.1 Characteristics of Speech

To ensure good features it is necessary to know the properties of speech. According to Stevens [3], there are some properties that set speech apart from other signals.

- The short-term power spectrum always has "peaks" and "valleys." These peaks in the power spectrum arise from the peaks observed in the vocal tract transfer function and correspond to the formants or vocal resonances that are so prominent in vowel and vowel-like sounds.
- The presence of up and down fluctuations in amplitude as a function of time. These variations in amplitude correspond to the alternation of consonants and vowels occurring in syllabic-like units roughly every 200–300 ms.
- The short-term spectrum changes over time. The peaks and valleys of the power spectrum change. Some changes occur very rapidly, such as the formant transitions of stop consonants, whereas other changes are more gradual, like formant motions of semivowels and diphthongs.

According to Stevens [3], speech sounds have these three general attributes and other sounds do not, and it is these attributes that distinguish speech sounds from other complex nonspeech sounds. It should also be mentioned that besides the differences just discussed, other marked differences exist in the manner in which speech and nonspeech signals are processed (that is, encoded, recognized, and identified) by human listeners. Human

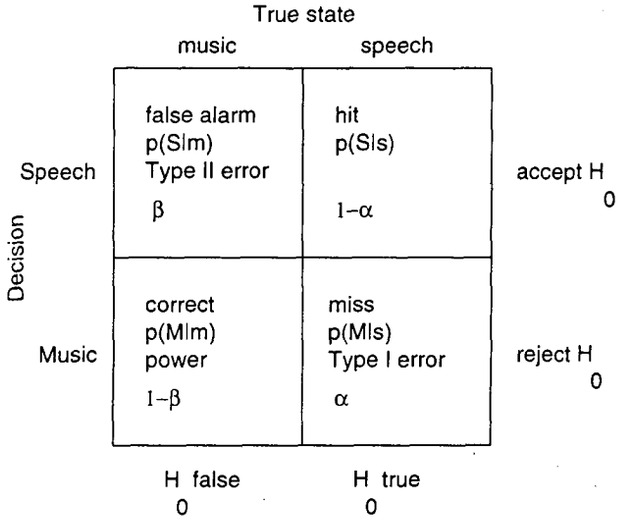


Fig. 3. Table illustrating two correct and two incorrect decisions possible when deciding whether to reject H_0.

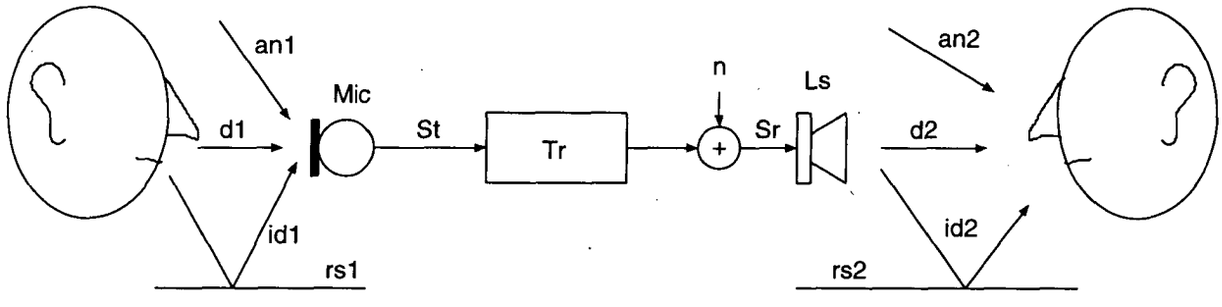


Fig. 1. Signal path from speaker to listener and some interfering signals.

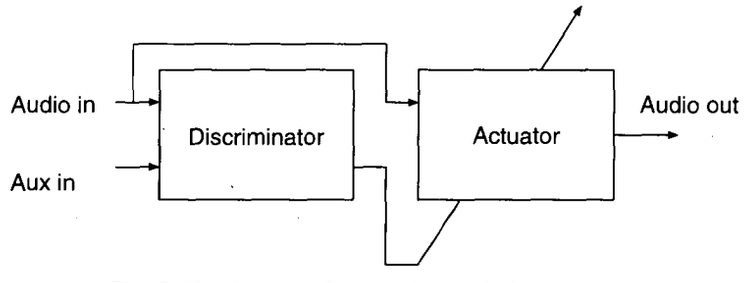


Fig. 2. Total system for speech–music improvement.

observers respond to speech signals as linguistic entities rather than simply as random auditory events [4].

2.2 Using Speech Properties

It is commonly accepted that the human ear "works" in the frequency domain as well as in the time domain. Thus candidates for feature extractors would be some kind of filter bank as used in [2], or some envelope detection of the audio signal has to be performed like that used in [5]. The latter is based on the differences in the envelope of music and speech signals; the frequency of abrupt drops in level is higher for speech than for music. A comparable method is described in [6], where a correct decision rate of about 86% was reached. A third approach could be to study statistical properties of the audio signal.

A more general approach is adopted by Sugiyama [7], who tries to recognize the language automatically. He does not mention the failure rate of the system when music is applied, but reports a recognition rate of 80% between 20 different languages. Another application for the speech–nonspeech discrimination is in the field of speech recognition, where it is very important to detect the boundaries of a word. A proposal by Kobatake et al. [8] is based on the detection of nonstationary segments by tracking the error in a linear prediction model (LPC). Rabiner and Schafer's proposal [9] is to study the number of zero crossings of the signal. A very simple and old method is to detect low levels in the difference signal (L–R) of stereo signals, the assumption being that voice is recorded monaurally. Other recent patents are by Kamiya and Ueda [10], who based their patent on the differences between the cepstrum coefficients of a frame and previous frames; and by Aarts [11], whose work is based on detecting features simultaneously in the time–frequency domain. A conceptual discussion on the processing of complex sound is given by Terhardt [12]. A broad but somewhat conceptual survey is presented in Bregman [13].

3 SYSTEM SETUP

An overall diagram of the speech–music discriminator is shown in Fig. 4. The system was implemented on a Motorola 56k digital signal processor. It consists of three parts: filtering and normalization (Section 3.1), a feature extractor (Section 3.2), and a fuzzy combiner (Section 3.3). The output of the discriminator is a signal $p'(s)$, which is an estimate of the probability that the input signal is a speech signal.

3.1 Filtering and Normalization

The filtering and normalization arrangement is shown in Fig. 5. The incoming audio signal is split into two paths. The upper path contains a bandpass filter (70–700 Hz) and a rectifier. The lower path contains a bathtub-shaped band-stop filter (130–1200 Hz) and a rectifier. The rectifiers are followed by low-pass filters (40 Hz) to obtain the mean power value of the filtered signal. The speech filter in the upper path will only pass speech in the frequency range of 70–700 Hz. If on the other hand music passes through the bandpass filter, the higher and lower frequencies (which are likely to be present in music) are eliminated so that the power contained in the music signal is reduced. The music filter in the lower path will pass speech with the same amplification level as the upper path. If music passes through, its lower and higher frequency components will be amplified so that the power contained in the music signal is boosted. When the input signal is speech, the signals in both paths are the same, resulting in a divider output of 1. If, on the contrary, a music signal is offered, the energy contained in the signal will be decreased or increased in the upper or lower path, respectively, resulting in the divider output being less than unity. Hence the divider output will be equal to zero when the input signal contains only frequencies below 70 Hz and above 700 Hz. Before the two signals are applied to the divider, the constants c_1 and c_2 are added. These constants are in the ratio of 1:2, but very small. So when the circuit is fed a normal input signal, the output will not change significantly because of the coefficients. However, when the input signal becomes very small, the coefficients will have an effect and the output will be about 0.5 and thus remain in the area between the high and low thresholds (see Fig. 6) so that no false decisions will be made. The upper path will be divided by the lower path, resulting in an amplitude-independent input signal. (Because of c_1

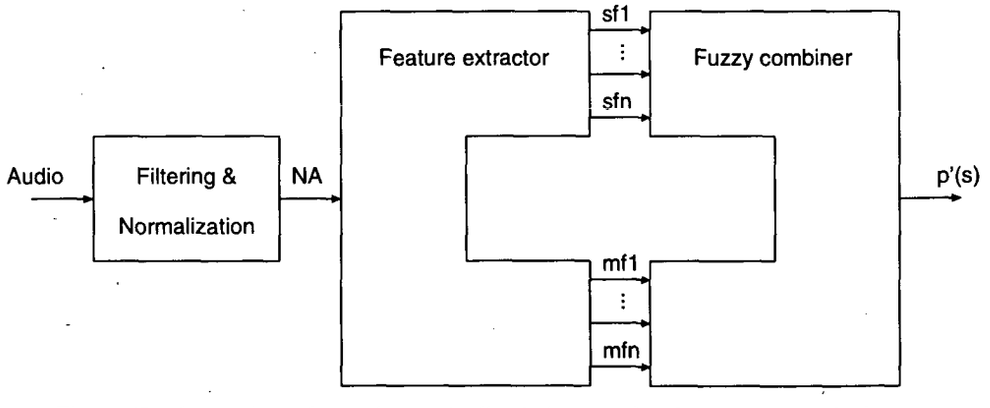


Fig. 4. Total overview of speech–music discriminator. sf, mf—speech, music features.

and c_2, the divider's output is, of course, not amplitude independent for very small input signals.) Finally, the divider's output will be fed into the feature extraction circuit.

3.2 Feature Extractor

The normalized audio signal (a typical signal is shown in Fig. 6) is analyzed for typical features occurring only in speech and music, giving speech and music features sf and mf, respectively. The extractor circuit is currently able to extract two features, which indicate if the audio signal is speechlike (referred to as sf_1 and sf_n in Fig. 4). The speech features can be best described in pseudo-Pascal (see Fig. 6 for an explanation of the terms):

```
if (tslope(n) < 100 ms) and (slope.up(n) = not(slope.up(n – 1))) and
  (tslope(n – 1) < 100 ms) and (slope.up(n – 1) = not(slope.up(n – 2))) and
  (tslope(n – 2) < 100 ms) and (slope.up(n – 2) = not(slope.up(n – 3))) and
  (tlastslope(n) < 700 ms) and
  (tlastslope(n – 1) < 700 ms)
then trigger sf1
if (45 ms < tbelowlowthreshold < 150 ms) then trigger sf2
```

Examination of the normalization circuit output signal has shown that a mean value is needed. The audio signal is not speech when the mean value lies below the mean threshold. The mean value is calculated over 100 samples in 1 second.

3.3 Fuzzy Combiner

Finally the speech and music features have to be combined in a single signal indicating an estimate of the probability that the input signal is a speech signal. The output $p'(s)$ is shown in Fig. 7.[2] A simple adder is used,

[2] The prime in $p'(s)$ denotes that it is an estimate for $p(s)$.

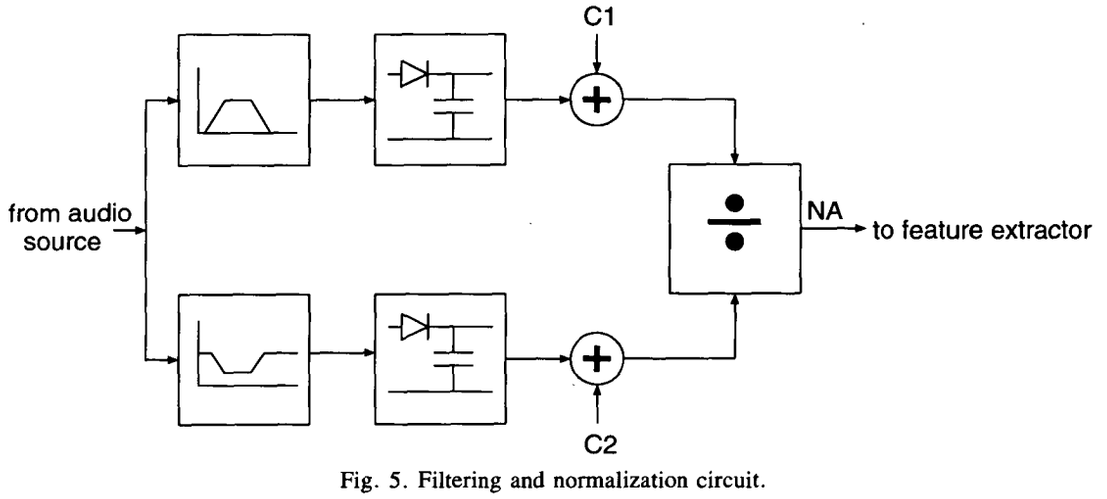


Fig. 5. Filtering and normalization circuit.

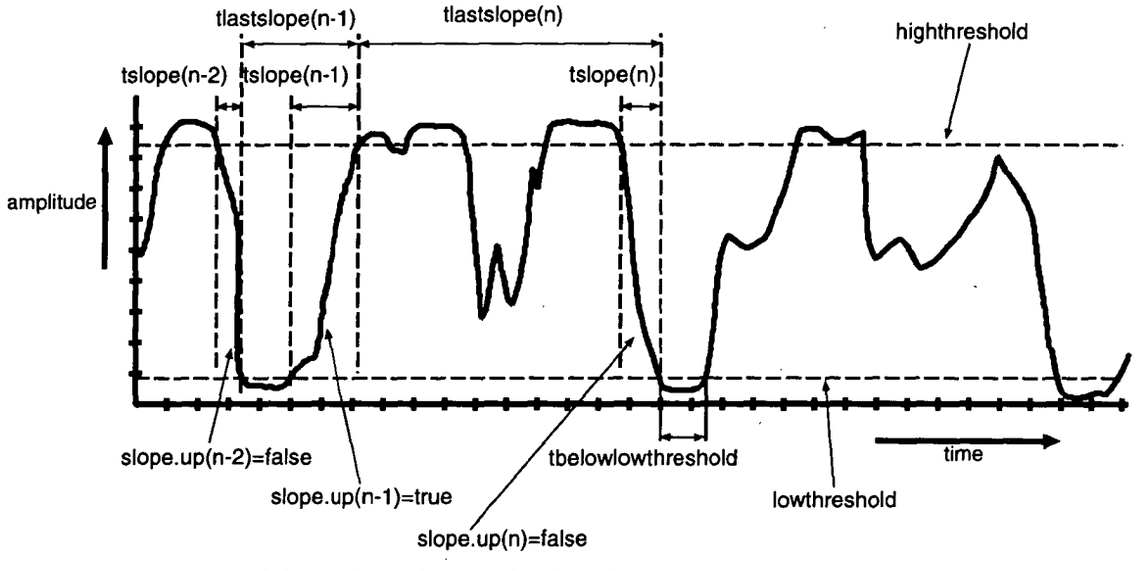


Fig. 6. Typical waveform produced by filtering and normalization section.

which increases the output by 0.5 each time a trigger (from sf_1 or sf_2) has occurred. The output is decreased if a music feature (mf) has occurred. If no feature occurs, then the output decreases at a rate of 0.1 per second, which gives the typical ramplike behavior indicated in Fig. 7. Finally the maximum output value is limited to unity and the minimum value to zero.

4 PERFORMANCE

To measure the performance of the discriminator, a variety of types of speech and music signals are used as test materials. The results of the measurements of $p(S|m)$, the false-alarm probability (defined as the number of triggers divided by the number of decisions), are presented in Table 1. Over 18 hours of tested music, the false-alarm probability is virtually zero, as shown in Table 1. To measure the hit rate, the number of speech features occurring in speech are counted. These results are presented in Table 2. As the table shows, there is considerable variability in the number of hits recorded by the various speakers, but in most cases the number of hits exceeds one per second, which might be sufficient for practical use.

5 CONCLUSIONS

It has been shown that a very reliable speech–music discriminator can be made, at least for "clean" signals, using a relatively simple real-time setup. The discriminator was tested for various languages spoken by persons of both sexes and for various types of music, and yields a false-alarm probability (inferring speech while the input is in fact music) of virtually zero, while in most cases the hit rate exceeds one per second.

6 REFERENCES

[1] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Lecture Notes, vol. 1 (Addison-Wesley, Reading, MA, 1991).

[2] N. S. Christensen, K.-E. Christensen, and H. Worm, "Classification of Music Using Neural Net," presented at the 92nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 40, p. 445 (1992 May), preprint 3296.

[3] K. N. Stevens, "Acoustic Correlates of Some

Table 1. Number of false speech features in music, false-alarm probability $p(S|m)$, and number of speech features (hits) per second.

Label	# sf	$p(S\|m)$	# sf s^{-1}
Dire Straits	0	0	0
Newman	3	29e-9	1.3e-3
A capella	14	23e-8	1.0e-2
McDonald	3	23e-9	1.0e-3
Collins	0	0	0
Madonna	3	51e-9	2.3e-3
Yello	7	67e-9	3.0e-3
Ferry	3	30e-9	1.3e-3
Yes	3	36e-9	1.6e-3
Rea	0	0	0
Norman	0	0	0
Mezzoforte	0	0	0
Mozart	0	0	0
1812	0	0	0
Pathétique	0	0	0
Johannes	0	0	0
Matthäus	0	0	0

Table 2. Number of speech features (hits) per second in speech signals.

Sex	Language	CD	# sf s^{-1}
F	English	SQAM	2.8
M	English	SQAM	1.1
F	French	SQAM	2.0
M	French	SQAM	0.7
F	German	SQAM	2.0
M	German	SQAM	1.0
F	English	Arch.	1.9
M	English	Arch.	1.2
F	Danish	Arch.	1.0
M	Danish	Arch.	0.3
M	English	Denon	1.1
M	Japanese	Denon	0.5
M	Dutch	Claus	0.3

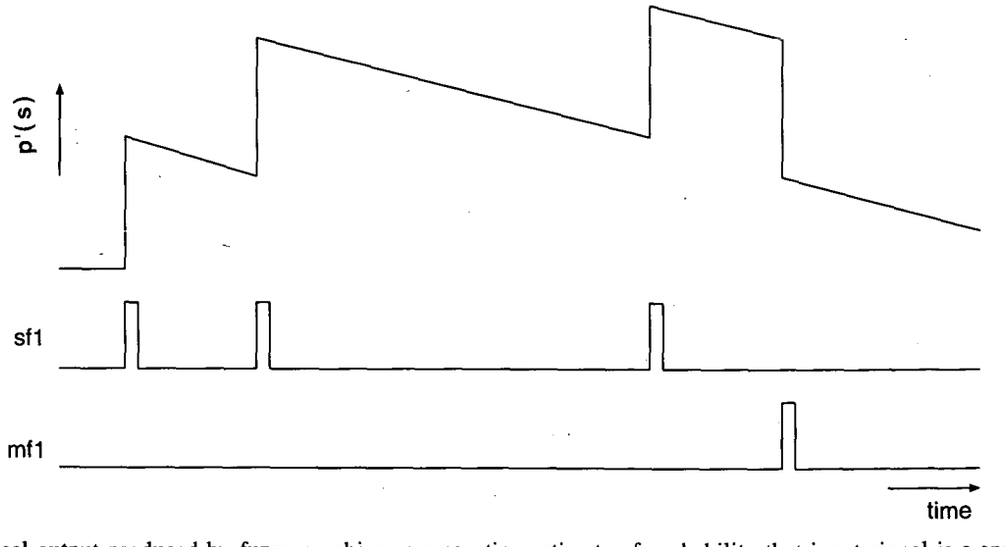


Fig. 7. Typical output produced by fuzzy combiner representing estimate of probability that input signal is a speech signal.

Phonetic Categories," *J. Acoust. Soc. Am.*, vol. 68, pp. 836–842 (1980 Sept.).

[4] D. B. Pisoni, "Some Comparisons of Speech vs. Nonspeech Signals," in *Auditory Processing of Complex Sounds*, W. A. Yost and C. S. Watson, Eds. (Lawrence Erlbaum Assoc., Hillsdale, NJ, 1987), pp. 247–256.

[5] E. Belger and H. Jakubowski, "Ein programmgesteuerter Musik-Sprache-Schalter (A Program Controlled Music–Speech Switch)," *Rundfunktechn. Mitt.*, vol. 12, no. 6, pp. 288–291 (1968).

[6] S. Okamura and K. Aoyama, "An Experimental Study of Energy Dips for Speech and Music," *Pattern Recogn.*, vol. 16, no. 2, pp. 163–166 (1983).

[7] M. Sugiyama, "Automatic Language Recognition Using Acoustic Features," in *Proc. ICASSP 91*, vol. 2 S2VLSI-U (1991 May), pp. 813–816.

[8] H. Kobatake, K. Tawa, and A. Ishida, "Speech/Nonspeech Discrimination for Speech Recognition System under Real Life Environments," in *Proc. IEEE/ICASSP*, vol. 1 (Glasgow, Scotland, 1989 May), pp. 365–368).

[9] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ, 1978).

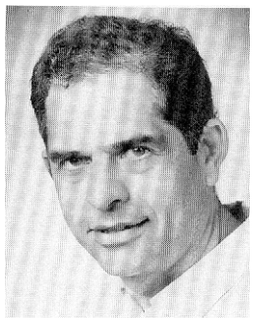
[10] S. Kamiya and T. Ueda, "Method of Distinguishing Voice from Noise," US patent 4,920,568 (1990).

[11] R. M. Aarts, "Device for Indicating a Probability that a Received Signal is a Speech Signal," US patent 5,878,391 (priority date BE9300775 (1993 July 26) (1999 Mar.).

[12] E. Terhardt, "Gestalt Principles and Music Perception," in *Auditory Processing of Complex Sounds*, W. A. Yost and C. S. Watson, Eds. (Lawrence Erlbaum Assoc., Hillsdale, NJ, 1987), pp. 157–166.

[13] A. S. Bregman, *Auditory Scene Analysis, The Perceptual Organization of Sound* (MIT Press, Cambridge, MA, 1990).

THE AUTHORS

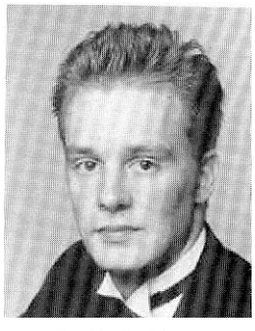


R. M. Aarts

Ronald Aarts was born in Amsterdam, The Netherlands, in 1956. He received the B.Sc. degree in electrical engineering in 1977 and the Ph.D. degree in 1994 from Delft University of Technology.

In 1977 he joined Philips Research Laboratories, Eindhoven, The Netherlands, in the Optics group. There he was engaged in research into servos and signal processing for use in both Video Long Play players and Compact Disc players. In 1984 he joined the Acoustics group and worked in the development of CAD tools and signal processing for loudspeaker systems. In 1994 he became a member of the DSP group and became engaged in the improvement of sound reproduction by exploiting DSP and psychoacoustical phenomena.

He has published a number of technical papers and reports and is the holder of several patents in the aforementioned fields. He was a member of the organizing committee and chairman for various conventions. He is a senior member of the IEEE, a fellow of the AES, the NAG (Dutch Acoustical Society), and the Acoustical Society of America. He is a past chairman of the Dutch Section of the AES.



R. T. Dekkers

Robert Toonen Dekkers was born in Eindhoven, The Netherlands, in 1969. He received a B.Sc. degree in electrical engineering from the College of Advanced Technology of Eindhoven in 1990. Since 1992 he has been employed by Philips Research Laboratories, Eindhoven, first in the Physical Acoustics group where he was engaged in research on stereo base widening, resulting in the Philips Incredible Surround feature. He later joined the Digital Signal Processing group.

In 1999 he joined a software company called PID, in Eindhoven. There he has been engaged in the development of Internet applications. His current activities involve the design of an advanced network system capable of automatic lighting control in large buildings.