Voice Characteristics of MARSEC Speakers

Eric Keller

LAIP - IMM - Lettres, Université de Lausanne, 1015 Lausanne, Switzerland,
eric.keller@imm.unil.ch

Abstract

An examination of Bark scale spectra of some 30'000 vowel nuclei from the Marsec corpus showed significant voice quality differences for sex and speech style. Of the two, speech style appeared to be the stronger predictor in multiple correlation analyses. Clustering further documented the effects of speech style, with hierarchical clustering grouping voice profiles of similar-sounding styles. Sports, news and market reports were at one end of the clustering tree, while fiction, poetry and dialogues occupied other distinctive areas on the tree. Further, a factor analysis of the Bark scale spectrum showed four areas of relatively independent variation for the common speech frequencies involved in transmitting voice quality, indicating that considerable parameter reduction is possible in the prediction of voice quality-related spectral variation. The results suggest that high-quality spectrally-based speech synthesis systems could profit from a systematic biasing of spectral profiles to convey vocal quality, particularly in the lower frequencies (up to 600 Hz) and in the higher frequencies of speech (above 1600 Hz).

1. Introduction

Speakers exhibit voice quality variation that reflects both their gender and the genre of thematic material under discussion. This variation must be modelled in high-quality speech synthesis systems. Spectral parameters in vowel nuclei relating to sex and speech style were investigated in the present study on the basis of 56 British English speakers from the Machine-Readable Spoken English Corpus (MARSEC[1]) [1, 2].

"Voice quality" necessarily designates "aspects of the final, audible waveform" here, rather than "acoustic and perceptual results of the state of the glottis", as the term is used more generally, e.g. [3, 4]. This widened terminology (*i.e.*, source *plus* filter) parallels the choice of spectral averaging techniques over inverse filtering approaches (such as the L-F model [5]) as the analysis tool in this study. Inverse filtering is difficult to perform in a large-scale study, since there are established difficulties in the automatic specification of initial parameters for inverse filters, significant source-filter interactions (primarily at weak glottal flows) neglected by inverse filtering, as well as the possibility of phase distortion with tape-recorded material [3, 6]. Inverse filtering is thus best performed manually and interactively, on selected vowel nuclei with normal or high intensity, best of all with material recorded directly to computer [3]. MARSEC, as a large-scale corpus with some 52'000 words, could thus not be analyzed with inverse filtering; also, because it was recorded on tape, phase distortions could not be excluded.

If spectral averaging techniques are more robust and easier to apply, they are not without methodological difficulties. Sufficient data must be available for each speaker to cancel out spectral modifications induced by pitch and formant structure in the spectrum. Material must be obtained in voiced portions of speech, preferably in the centre of vowel nuclei. Finally, intelligent data reduction must be performed for statistical evaluation. In the present study, these concerns were addressed by the development and use of an automatic vowel nucleus detector, by averaging a minimum number of sample points from each subject, and by the identification of spectral response regions.

MARSEC presents a number of advantages for this type of study. Subjects were relatively numerous and were all representatives of the same received-pronunciation (RP), *i.e.* the prestige/regionally neutral accent form of British English, all were practiced adult speakers, a good spectrum of speech styles was represented, and the recording quality was generally quite good, since corpus selections were either broadcasts from the BBC or course material produced by Open University. Furthermore, several speakers contributed extended speech material to the corpus, which will ultimately permit some intra-subject comparisons. Of relevance to studies of voice quality, a large, naturalistic corpus such as MARSEC permits to initiate the examination of the statistical effects of the finer shades of gender, attitude and thematic coloration of voice that have largely been left aside in studies bearing on the voice quality of emotion portrayals. Finally, the prosodics of the speech in MARSEC have been extensively examined elsewhere (see [2], [7]), which simplifies further examination of links between voice characteristics and prosodic features.

In this study, questions concerning both independent and dependent voice quality variables were investigated. With respect to predictors, the effects of sex, speech style and individual characteristics were examined. Concerning dependent variables, spectral ranges interacting with the independent variables were delimited. *A priori*, sex, speech style and subject identity were all expected to be systematic sources of spectral profile variation, because to the ear, women sound systematically different than men, sports reports sound different than religious broadcasts, and individuals all sound different from each other to some degree. Also, since adjoining Bark bands are known to be fairly strongly inter-correlated (see e.g. [9]), it was expected that spectral regions larger than Bark bands could serve as distinctive carriers of voice quality. If successful, profiles developed with respect to the parameters examined here might serve to induce appropriate voice quality modifications in high-quality speech synthesis systems.

[1] MARSEC is available from www.rdg.ac.uk/AcaDepts/ll/speechlab/marsec.

2. Method

2.1. Corpus and Subjects

MARSEC incorporates material from 65 speakers, transcribed textually and prosodically, with word-level timing alignment. By design, eleven speech styles were incorporated entitled "commentary, news, lectures I and II, religious broadcasts, magazine reports, fiction, poetry, dialogue, propaganda and miscellaneous". However, upon listening to the entire corpus and performing a number of perceptual and thematic comparisons, a somewhat different and not entirely overlapping prosodic style categorisation was arrived at. Styles distinguished on the basis of (a) the author's prosodic perception and (b) thematic categorisation were as follows (MARSEC categories in parentheses) [1]:

Report	short reports on various topics from correspondents in the UK and abroad (A Commentary, B News [correspondent reports], K Propaganda)
News	"anchor" news and weather, program announcements (B News [excluding reports], M02-04, M07-09 Miscellaneous)
Lecture	types I & II, for the general public and Open University students respectively (C Lecture type I, D Lecture type II)
Citation	citations of authors, read aloud (found in the first of D Lectures type II)
Religious	religious speech, including liturgy (E Religious broadcasts)
Market	market and financial news (F0101-0308 Magazine-style reporting)
Fiction	general fiction aimed at adults or children (G Fiction, M01 Miscellaneous)
Sports	yearly review of sports events (F04 Magazine-style reporting, J01 Dialogue)
Dialogue	ESL dialogues in the form of radio plays (J02-06 dialogue)
Address	formal, public addresses (M05-06 Miscellaneous)

MARSEC material was recorded during the early and mid 1980s at broadcast studio quality, with the exception of a number of correspondent reports transmitted from abroad by telephone. These suffered from varying degrees of audio channel reduction, none serious, but of sufficient importance to a study of vowel quality. The passages marked as "dialogue" were spoken for a public of advanced ESL students and are somewhat slowed and hyperarticulated (particularly J06). All remaining material is appropriate to native-English BBC or Open University audiences.

For the present study, signals from the corpus were subdivided at the major tone group ("sentence") boundaries, marked by double bars in the prosodic transcription. Each sentence was labelled for (a) speaker (b) sex of speaker and (c) speech style. Six sentences with speaker overlap were eliminated. Of the total of 65 subjects, 56 produced more than 25 vowels and were used for this study.

2.2. Vowel Nucleus Identification and Spectral Profile Analysis

2.2.1. Vowel nucleus detector

Simple long-term spectra are contaminated by spectral measures from voiceless and silent speech portions, which are relatively uninformative about vocal quality. One approach has thus been to use spectral profiles from voiced portions of speech only [9]. In the present study, a vowel nucleus detector was developed using a statistical model of the effects of vowels and pauses on a 22 Bark-band filter[2]. To create bark band spectra, log FFT magnitude spectra were derived from the acoustic waveform at 500 Hz. With 16 kHz signals, 1024-pt hamming windows with pre-emphasis were calculated. Bark scale mean magnitude values were obtained from each 8 kHz magnitude spectrum, and skew in the distribution of the values was approximately corrected with a square root operation. Subsequent to normalisation, values were more or less normally distributed (mean kurtosis: 1.15, mean skewness: -0.86). These values served as inputs to a linear regression model for predicting the presence of vowels in 87 sentences with manually verified segmentation, taken from a speaker in the corpus (BP from BBC-4 News).

Since initial results showed frequent confusions between silent periods and vowel nuclei, a second model for predicting silent periods was created in similar fashion. Vowel nuclei were only accepted for the study when the prediction for silent periods descended below a given (fairly severe, optimised) limit. With this second detector, a Pearson correlation of 0.889 was obtained between 1067 manually identified and automatically estimated vowel spectra ($p<0.01$). Errors included 69 spectra from voiced segments, 17 from silent periods and 34 (or about 3.2%) from unvoiced spectra. A total of 30'453 vowel nuclei from 61 subjects were identified in this manner in the corpus.

2.2.2. Spectral profiles and standardisation

Spectral slices were obtained by centring 1024-pt FFT windows in the corpus' 16 kHz signals at points indicated by the vowel nucleus detector. As before, spectra were bark band averaged and square-root adjusted for skew. Considerable variability was observed for Bark band responses. This was likely due to three factors: (1) overall *amplitude differences* resulting from differences in recording volume (bit filling) and in voice volume, affecting the entire spectrum (see Figure 2), (2) *local phonetic factors*, such as the placement of harmonics and formants in the spectrum, and (3) *stable* as well as *variable individual voice characteristics*. Since it is the last factor group that we are most interested in, the effects of the first two factors must either be kept constant or be suppressed by averaging techniques. Averaging was chosen, since it was difficult to control overall amplitude or local phonetic factors. Individual spectral profiles were thus z-scored with respect to the grand individual mean and standard deviation, *i.e.*, the mean and sd. obtained for each subject over all Bark bands and all vowels (*z-score:* distance from the mean normalized by the standard deviation of the distribution). In addition, it was examined how many spectra were needed to obtain stable mean profiles. Approximate

[1] The detailed list is available from the author.

[2] The rounded Bark scale limits were as follows (Hz): 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, and 8000 (Nyqvist limit). Traunmüller's formula ($B = 26.81/(1+(1960/f)) - 0.53$) appears to give a fairly close approximation of empirical Bark values in the 0-8 kHz range (see Paul Carter's [University of York] review page for various formulas to calculate Bark band values: http://www-users.york.ac.uk/~pgc104/phonlink/bark.html.)

stability was observed for averages of 25+ spectra, and subjects providing fewer than 25 spectra were eliminated. This left z-scored spectra for 30'401 vowel nuclei from 56 subjects available for analysis. The distribution of these spectra was, by style of speech, address 347, citation 117, dialogue 1'206, fiction 729, lecture 1 213, lecture 2 642, market 2669, news 989, poetry 1'634, religious 653, report 7'195 and sports 14'007 vowel nuclei, and by sex: male 24'122 and female 6'279 vowel nuclei.

2.3. Analysis and Results

2.3.1. Areas of common spectral response

To identify spectral regions involved in voice quality variation, a principal component analysis was performed on the data set. Factor loadings in the same (positive or negative-going) direction indicate correlated variation. Since inter-correlated spectral magnitudes have been reported for adjoining filter bands, we expected one or more important factors to show similar loadings for two or several adjoining Bark bands. That was indeed the observed pattern. Five factors with Eigenvalues > 1.0 explained a cumulative 41.8% of total variance (see Figure 1). The first three showed fairly strong, same-directional factor loadings for adjoining bark bands. Factor 1 (9.2% of variance) subdivided the spectrum into just two regions by combining low-frequency variation below about 600 Hz (Bark 5.7[1]) with opposite-going variation above that frequency. Factor 2 (3.4% of variance) provided a tri-partite subdivision, with strong same-directional variation below about 1080 Hz (Bark 9) and above 5300 Hz (Bark 19). Factor 3 (2.6% of variance) showed opposite-going variation between a low-frequency (below 770 Hz, Bark 7), a mid-range (770–3150 Hz, Bark 7-16) and a high-frequency component (above 3150 Hz, Bark 16). Factor 4 (1.3% of variance) again showed activity counter-correlated between the mid band (500 Hz/Bark 5 – 1400 Hz/Bark 10.6) and a high-and-low band (below 500 Hz/Bark 4.9 and above 1400 Hz/Bark 10.6). Factor 5 (1.1% of variation) reflected further complex, but minor deviation.

Based on zero-crossings and major deflection points of the first three factors, four major regions of fairly independent spectral magnitude activity were differentiated in the factor loadings (albeit with a bit of arbitrariness). These regions can be said to show substantial same-directional behaviour. *Region 1*: 0-600 Hz (Bark 0-5.7), *Region 2*: 600-1600 Hz (Bark 5.7-11.5), *Region 3*: 1600 – 5000 Hz (Bark 11.5-18.7), and *Region 4*: Bark bands 5000 – 8000 Hz (Bark 18.7-21.3). Subsequently, average filter band responses were calculated for these four spectral regions, and the statistical procedure in the next step used averaged values from these four spectral regions as dependent variables (even though the figures, for reasons of explicitness, show spectral profiles with Bark bands).

2.3.2. The effects of sex and speech style

Figures 3 and 4 show standardized average profiles for sex and style of speech. Although visually, the differences appear minor, one-way ANOVAs indicated significant differences for nearly all factors and all spectral regions. One-way ANOVAs for 2 factors * 4 regions for sex showed $F_{(1, 30401)}$ = 0.000, 178.304, 229.419 and 645.060 for regions 1-4 respectively, significant at p=.989, .000, .000, .000. One-way ANOVAs for 11 factors * 4 regions for style showed $F_{(1, 30401)}$ = 1530.19, 77.48, 387.68, 327.05 for regions 1-4 respectively, all significant at p = .000. The only non-significant difference was thus for sex in spectral region 1[2].

Multiple regression analyses were run to examine if these differences could be used for a predictive model, and if so, which predictor levels would be attained. Exceptionally strong interactions were in evidence between *subject identity* and both *sex* and *speech style*, threatening the generality of the model. The factor *subject* was therefore excluded, leaving the general model Intercept + STYLE + SEX + STYLE*SEX for testing in the four spectral regions. Wilk's Lambda was significant at p = .000 for all three predictors (F_{style} = 262.610, F_{sex} = 56.106, $F_{style*sex}$ = 176.130), and the model provided correlations of r = 0.647, 0.210, 0.421 and 0.366 respectively for the four spectral regions (average: r = 0.411). An analogous multiple regression on the 22 Bark bands gave similar values with p-values significant at .000 for F_{style} = 294.537, F_{sex} = 38.674, $F_{style*sex}$ = 220.857) and a similar average correlation (r = 0.402[3]).

These results can be summarised by saying that sex and speech style have significant effects on the overall spectral profile of vowels, particularly so in the first and third spectral regions, and that of the two examined predictors, style appears to be a better predictor than sex in the standardized data. The average correlations show that the strongest predictions are made for the first spectral region. Also, the similarity in the prediction of spectral magnitude values in spectral regions and in Bark bands provides an initial validation of the subdivision of the spectral domain performed in the previous step.

2.3.3. Underlying factors

To suggest parameters underlying similarities between spectral profiles, a clustering analysis was performed using the full bark data. Hierarchical clustering again works on the principle of correlated variation, and it calculates relative proximity between predictor values on the basis of the proximity between successively larger groups of variables. The results appear in Figure 5. The link between speech style and spectral profile is confirmed. Sports, news and market reports were found predominantly toward one end of the dendrogram, religious speech, dialogues and telephone-transmitted reports were found toward the other end, and fiction and poetry tended to cluster in the middle. It is interesting to note that the telephone-transmitted reports all clustered toward one end of the spectrum. Further, women's voices were interspersed with men's, indicating that spectral similarity between speakers discussing similar topics was again of greater influence than similarity relating to gender.

To the ear, sportscasters' voices on the MARSEC sounded relatively tense, while fiction readings sounded relatively lax. This suggested that the well-documented tense-lax continuum [6, 8] may well form an important underlying

[1] Bark values here are calculated with Traunmüllers formula.

[2] A comparison of figures 2 and 3 shows that this may be partly an effect of the standardization procedure.

[3] The slightly lower value for Bark bands is due to different average weighting produced by 22 Bark bands *vs.* the 4 spectral regions.

parameter of these clustering patterns, a hypothesis that could be investigated by examining inverse filtered source files from these speakers. Another possibility is that the spectral differences may be related to differences in speech rate among the various speakers. Research on the contribution of these possible factors is planned for later.

3. Discussion

The results of this study indicate that spectral profiles are clearly related to sex and speech style. Of the two, speech style appears to have stronger predictive value. Further, a factor analysis of Bark scale spectrum showed four areas of relatively independent variation for the common speech frequencies involved in transmitting voice quality, suggesting that a possible parameter reduction from 22 to four spectral regions for voice quality-related spectral variation.

It is interesting to note that the region permitting the least amount of prediction for voice quality variation is also the region where the strongest linguistically relevant formant information is found, i.e., the region between 600 and 1600 Hz. It may well be that the paralinguistic aspects of speech, *i.e.* vocal quality and prosodics, are more strongly represented in the regions where linguistically relevant formant information is less in evidence.

The current findings could be tested systematically by varying spectral profile biasing of output spectra with spectrally-controlled speech synthesis, such as HNM or formant-based systems.

4. Acknowledgements

Mr. L. Wiget is thanked for his diligent work with manual segmentation. Thanks are also extended to Dr. B. Zellner-Keller, Mr. L. Wiget, Prof. J. Schwyter and Prof. J. Schoentgen for their suggestions on a previous version of this manuscript. This research is supported by a Swiss OFES grant in support of work performed under COST 277.

5. References

[1] Knowles, G., Williams, B. & Taylor, L. *A Corpus of Formal British English Speech.* Addison Wesley Longman. 1996.

[2] Knowles, G., Wichmann, A., & Alderson, P. *Working with Speech.* Addison Wesley Longman. 1996.

[3] Ní Chasaide, A. & Gobl, C. Voice source variation. In W.J. Hardcastle and J. Laver, *The Handbook of Phonetic Sciences* (pp. 427-461). Oxford: Blackwell. 1997.

[4] Epstein, M.A. Voice Quality and Prosody in English. Ph.D. dissertation, UCLA. 2002. Available from http://www.linguistics.ucla.edu/people/grads/melissa.

[5] Fant, G., Liljencrants, J., & Lin, Q. A four parameter model of glottal flow. *STL-QPSR, Vol. 4, 1-13. 1985.*

[6] Strik, H., Cranen, B. & Boves, L. (1993). Fitting a LF-model to inverse filter signals. EUROSPEECH-93, Berlin, Vol.1, pp. 103-106.

[7] Arnfield, S. C. Prosody and Syntax in Corpus Based Analysis of Spoken English. Ph.D. thesis, University of Leeds, UK. (1994). Available from http://www.linguistics.rdg.ac.uk/staff/Simon.Arnfield/phd.html.

[8] Ladefoged, P. *Preliminaries to Linguistic Phonetics*. Chicago: University of Chicago Press. 1971.

[9] Banse, R., & Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70, 3.614-636.*

6. Annex

Here are z-score values for the 22 Bark bands, averaged over the 56 subjects of the corpus: -0.558658098, 0.323945954, 0.949984862, 1.022225339, 1.013304395, 1.136413577, 1.279580391, 1.171089663, 1.182849502, 0.99992414, 0.921878945, 0.791521071, 0.765144033, 0.711587209, 0.644032813, 0.600522558, 0.516713255, 0.440050227, 0.485886821, 0.264329132, 0.068472948, -0.360613146.

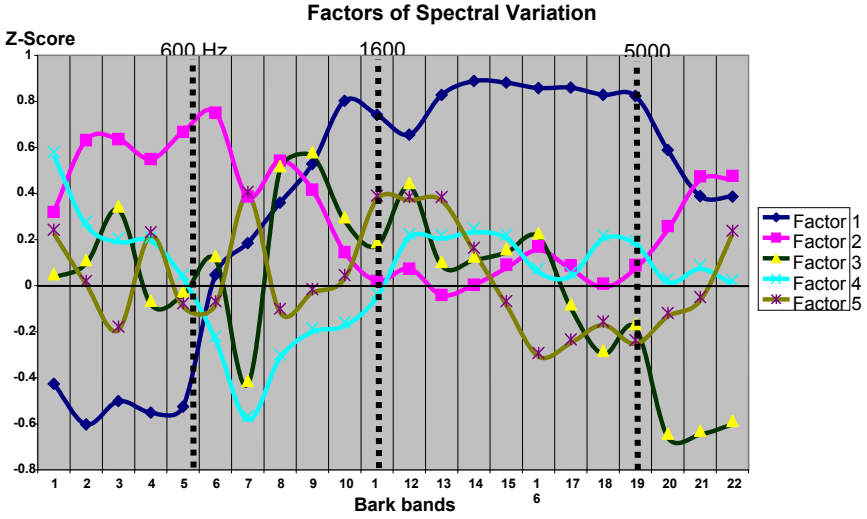

Figure 1. Five factors explained abut 42% of the variation in voice nucleus spectral profiles (based on all 30'401 spectra). Four major regions of roughly same-directional variation were identified by this analysis, delimited by 600, 1600, 5000 and 8000 Hz (see text).

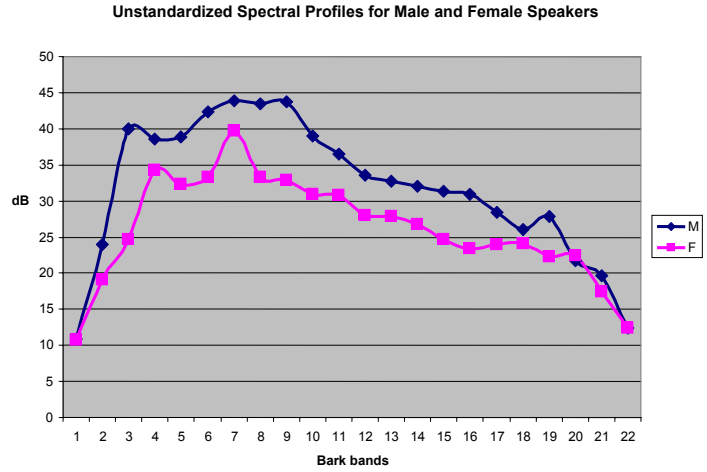

Figure 2. Spectral variation as a function of sex, unstandardized. Female speakers show globally lower volume than male speakers in this corpus.

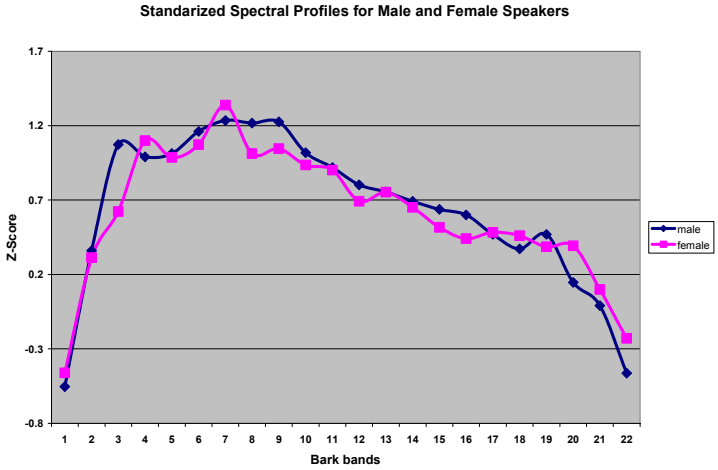

Figure 3. Spectral variation as a function of sex, standardized. The only major differences occur around Bark bands 3 and 8-9. The statistical analyses of this study were performed on z-score standardized data (see text).

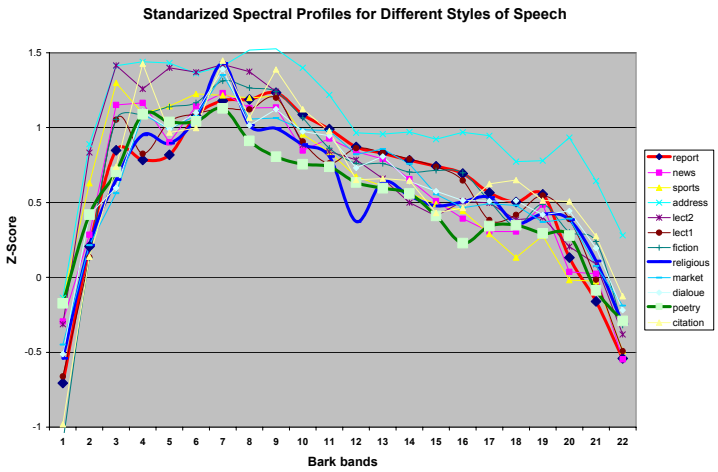

Figure 4. Spectral variation as a function of style of speech, standardized. Reports, religious speech and poetry occupy regionalised areas in the cluster analysis shown in Figure 5 and are marked with strong lines in this figure.

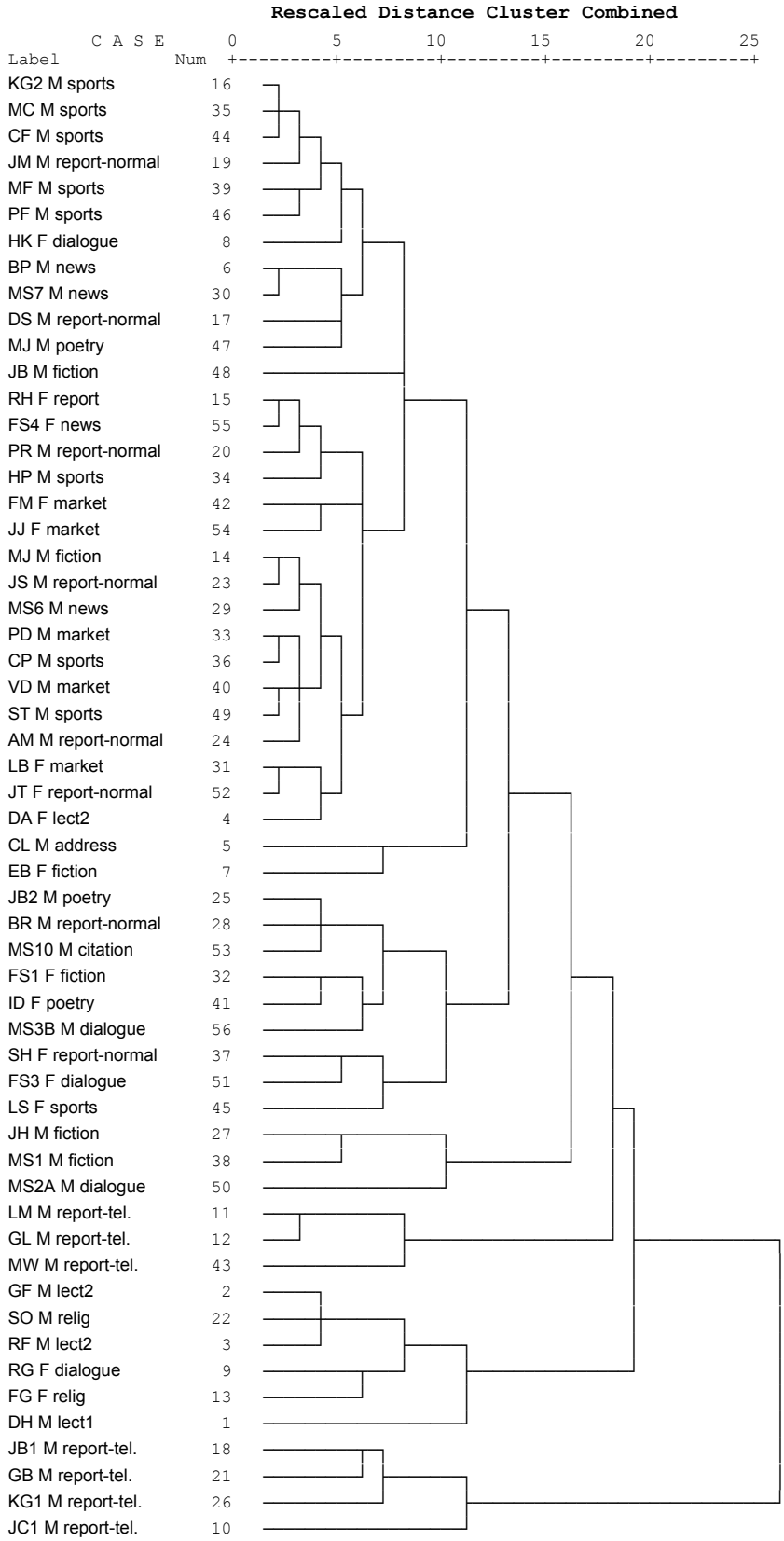

Figure 5. Cluster dendrogram obtained from the spectral profiles with the SPSS hierarchical clustering method for average linking. One notices the clustering of sports, news and market reports at one end, and the prominent presence of religious speech, dialogues and telephone-transmitted reports at the other end of the tree. Fiction and poetry tend to cluster in the middle.