HAL
archives-ouvertes.fr

Automatic intelligibility measures applied to speech signals simulating age-related hearing loss

Lionel Fontan, Jérôme Farinas, Isabelle Ferrané, Julien Pinquier, Xavier Aumont

▶ To cite this version:

HAL Id: hal-01343047
https://hal.archives-ouvertes.fr/hal-01343047

Submitted on 7 Jul 2016

Automatic intelligibility measures applied to speech signals simulating age-related hearing loss

Lionel Fontan[1], Jérôme Farinas[1], Isabelle Ferrané[1], Julien Pinquier[1], Xavier Aumont[2,3]

[1]Université de Toulouse; UPS; IRIT; Toulouse, France
[2]Université de Toulouse; UT2J; LETRA; Toulouse, France
[3]Archean Technologies; Montauban, France

{fontan, jfarinas, ferrane, pinquier}@irit.fr, xaumont@archean.fr

Abstract

This research work forms the first part of a long-term project designed to provide a framework for facilitating hearing aids tuning. The present study focuses on the setting up of automatic measures of speech intelligibility for the recognition of isolated words and sentences. Both materials were degraded in order to simulate presbycusis effects on speech perception. Automatic measures based on an Automatic Speech Recognition (ASR) system were applied to an audio corpus simulating the effects of presbycusis at nine severity stages. The results are compared to reference intelligibility scores collected from 60 French listeners. The aim of this system being to produce measures as close as possible to human behaviour, good performances were achieved since strong correlations between subjective and objective scores are observed.

Index Terms: Presbycusis simulation, Speech intelligibility metric, Automatic speech recognition.

1. Introduction

Age-related hearing loss – or *presbycusis* – refers to a progressive, bilateral and rather symmetrical sensorineural hearing loss occurring in people aged 50 and over. Its consequences on speech perception may greatly impact listeners' ability to communicate, especially in noisy environments [1]. In order to measure this effect in individuals and to tune hearing aids accordingly, speech intelligibility tests are often used as a complement to pure tone audiometry. They usually consist in the audio presentation of lists of words and/or sentences; the listener has to repeat or write down what he/she heard. Precision scores (percentage of correct words/sentences) are eventually computed [2]. Speech intelligibility tests have three main disadvantages:

1. They are rather tedious. In addition to a proper audio equipment, they require the presence of one or several jury members for the validation of the answers and the computation of scores (which often remains manual).

2. The linguistic material is often limited to isolated words lists. Recent studies suggest that word intelligibility scores resulting from such tasks are poorly correlated with the performances of listeners processing sentences in a particular context [3, 4].

3. In the case of presbycusis, the interpretation of speech reception scores is not straightforward for aged subjects who may suffer from cognitive troubles also interfering with their ability to understand speech.

In order to cope with these issues, the long-term goal of this research work is to create a framework for providing objective and automatic measures ranging from words intelligibility to sentences comprehensibility scores, by means of Automatic Speech Recognition (ASR). Similar approaches have been conducted in order to evaluate motor speech production disorders and their consequences on communication [5, 6]. To the best of our knowledge, this is the first study dealing with the prediction of intelligibility scores within the case of a hearing trouble.

This present work concentrates on the comparison of subjective and automatic measures of speech intelligibility during recognition of a) isolated words and b) sentences in French, both materials being degraded in order to simulate presbycusis effects at nine severity stages (see section 2). This material was used to collect reference intelligibility scores from 60 French listeners (see section 3). In order to generate automatic scores, an ASR system was tuned (see section 4).

The main motivation of our study, and its novelty, is not to improve automatic speech recognition results but to stay close from the tendencies observed in the human scores. To reach this goal, different ways to model word sequences were used. Results, in the form of descriptive curves and correlation scores, are presented in section 6 and discussed in section 7.

2. Simulation of presbycusis

Two main methods are used to simulate presbycusis [7]. One is to use steady noise such as white noise to mask speech and therefore reduce its audibility; however this method is not applicable to severe losses because an unbearable level of noise presentation would be required [1]. The other method consists in reproducing the three main typical effects of presbycusis:

- reduced audibility – especially in high frequencies. To simulate this effect, thresholds in several frequency bands are elevated according to the subject's audiogram.

- reduced frequency selectivity can be simulated by a spectral smearing algorithm [8].

- loudness recruitment. Due to a reduction of the intensity dynamic in the inner ear, the subject perceives intensity changes in an excessive way and becomes progressively intolerant to loud noises. Raising the signal envelope can be used to simulate loudness recruitment [9].

To create a speech corpus simulating the effects of presbycusis on speech perception, three speakers (a 12 year-old female child, a 47 year-old female adult and a 46 male adult) were recorded in an audiometric booth with an omnidirectional

Sennheiser MD46 microphone and a TASCAM DM-3200 mixing console. Each speaker read aloud 70 words and 70 sentences extracted from lists widely used by French audiologists: disyllabic word lists originating from [10] and French version of the Hearing in Noise Test – HINT [11]. For each list three listeners adjusted the level of each item (word or sentence) by comparing it with a reference item; at last the mean gain set by the three listeners was applied on the original audio files.

The algorithms, described in [12], simulate reduced audibility, frequency selectivity as well as loudness recruitment. These processes depend on the severity of the trouble to be simulated, which is represented by an audiogram indicating the hearing loss dB values for 15 frequencies ranging from 125 Hz to 16 kHz. In order to represent different severity grades corresponding to typical losses in presbycusis, nine typical audiogram values have been calculated by averaging audiometric data collected by [13] on 3753 subjects (see table 1).

Speech stimuli were processed by applying filters corresponding to the nine audiogram values. To limit the number of variables, and since loudness recruitment evolution can vary greatly in individuals [1], its simulation was set to only one condition. To this purpose the envelops of speech signals have been raised to the power of 2, which corresponds to the simulation of a moderate hearing loss. In [12] the spectral smearing algorithm depends on the severity grade to be simulated; to this end the severity grade is defined as a function of the subject's mean loss in dB for frequencies ranging from 2 kHz to 8 kHz. The nine degradation algorithms have been applied to the 140 original speech stimuli, resulting in 1260 final stimuli.

3. Subjective measures

Sixty listeners aged from 18 to 30 years old participated in intelligibility tests prior to this study [14, 15]. All participants were native French speakers, and did not suffer from any hearing loss superior to 15 dB, on average, between 2 kHz and 8 kHz. Thirty listeners participated in a word intelligibility test and the other half participated in a sentence intelligibility test (i.e., repetition of disyllabic words and sentences materials presented in previous section). For both tests, listeners were placed in an audiometric booth, sitting at approximately 3 feet away from two Tannoy Precision 6D loudspeakers. Participants were first familiarized with the procedure by listening a set of 10 items, and were then asked to repeat 60 stimuli: 54 stimuli degraded following the nine conditions exposed in section 2 and 6 original stimuli. The degradation conditions were distributed among the participants, and stimuli were presented in a random order.

Participants' responses have been recorded and transcribed, in order to be later compared with the output of the ASR systems described in next section.

4. ASR systems description

All automatic measures were obtained from ASR systems based on Sphinx-3 [16] distributed by Carnegie Mellon University.

4.1. Acoustic models and speaker adaptation

Acoustic models used in this study have been made available by the *Laboratoire d'Informatique de l'Université du Maine* (*LIUM*) for speech recognition in French language [17, 18]. They were trained on the basis of French radio broadcasts [19], under the form of continuous models consisting of 5725 context-dependent states (senones) with 22 gaussian mixtures

per state. Acoustic models include 35 phones as well as five kinds of pauses, and are meant to process 16 kHz speech samples using a PLP feature extraction [20].

In the training corpus male speakers were more represented than female and child speakers [21]). The ASR system was therefore likely to present better results with speech stimuli recorded by the male speaker. To check this hypothesis, the system performances were first observed with a lexicon restricted to the 60 dissyllabic words included in the lists. As expected, stimuli recorded by the male speaker are better recognized than stimuli recorded by the two other speakers (figure 1).
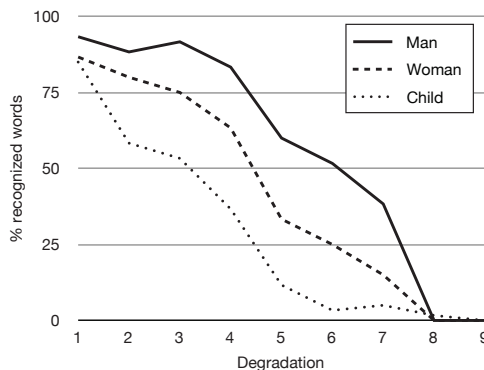

Figure 1: *Word recognition scores per speaker using the LIUM ASR and a vocabulary limited to the words from the stimuli*

As a consequence, the first step of this present work was to adapt the system to the speakers. To this end the vocal tract length normalization (VTLN) technique [22] was used. The VTLN supposes a linear relationship between the vocal tract length of the speaker and formant areas; following this idea, numerous research works have been devoted to the determination of the best frequency warping factor λ maximizing the likelihood of the phones produced by a speaker:

$$\lambda = arg\,max\,P(O|X, \lambda_k) \qquad (1)$$

X being the observation and k the index of the k-th frequency warping factor considered. The inverse linear function ($y = \frac{x}{\lambda}$) implemented in Sphinx-3 was used. Different values for λ were tested and the corresponding recognition scores were observed. Figure 2 presents word recognition results for isolated disyllabic words, with λ varying from 0.4 to 4.0. Polynomial regression equations have been calculated in order to determine the best warping factor for each speaker. For the male speaker $\lambda = 1.0$ (i.e., no adaptation is necessary) and for female and child speakers the optimal λ values are respectively 1.84 and 2.24.

4.2. Lexicon and Language models

In order to generate recognition scores for both isolated words and sentences, the trigram language model set up by the LIUM [17, 18] has been used. In addition to this baseline model, two other models have been specifically generated:

- for isolated words lists, a bigram language model was created to take into account the lexical and phonological characteristics of the target stimuli and extend the lexical coverage of the word hypotheses. Each stimulus consisting of a definite article and a noun beginning with

Table 1: Degradation level, theoretical age (years) and mean hearing loss values (dB), with associated severity grades

Degradation	Th. Age	125Hz	250Hz	500Hz	750Hz	1kHz	1.5kHz	2kHz	3kHz	4kHz	6kHz	8kHz	Severity
1	60	12	12	12	12	13	16	20	29	36	42	45	Mild
2	66.25	15	15	15	15	16	20	25	35	43	49	55	Moderate
3	72.5	18	18	19	20	21	26	31	42	50	57	64	Moderate
4	78.75	22	22	24	25	27	32	37	48	56	64	71	Moderate
5	85	27	28	30	32	34	40	45	54	62	70	77	Severe
6	91.25	32	34	37	40	43	48	53	61	68	77	81	Severe
7	97.5	39	41	46	50	53	59	62	67	73	83	84	Severe
8	103.75	46	50	56	61	65	70	72	73	78	88	84	Severe
9	110	55	60	68	74	78	83	83	80	82	90	83	Severe
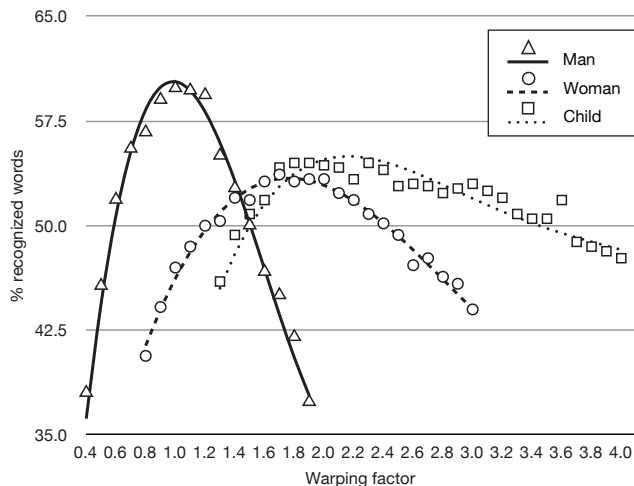
Figure 2: *Mean recognition scores as a function of VTLN warping factor*

Table 2: Codes for language models used in this study

Language model	Code
LIUM system with 62k words lexicon	LIUM-62K
LIUM system with restricted lexicon	LIUM-RES-LEX
Bigrams (Det + N) with Lexique 3.8 freq.	BIG-DET-N
Finite-State Grammar	FSG

a consonant (e.g. *le vacher*, [l@vaSe]), a list of 15146 masculine nouns was used. To reflect the frequency of these forms, the frequencies values defined by [23] and available in the database Lexique 3.8[1] were also used.

- for sentence materials, the LIUM system was used with its 62k-word lexicon and with a reduced lexicon limited to the words included in the 60 sentences (namely 366 words). A finite-state grammar was also defined to represent all the lexico-syntactic structures present in the sentences. These structures are not syntactically complex: each sentence is constituted by a single assertive clause describing an action or a state (e.g. *le chien dormait dehors*, "The dog was sleeping outside").

To clarify the results, all the system configurations used in this study are associated with codes given in table 2.

5. Subjective and automatic scoring

Concerning subjective measures only percent of correct words were considered. As usually observed in literature, words repeated with at least one phoneme different from the initial stimulus were marked as incorrect.

On the other hand, several automatic measures were considered. The first measure is equivalent to the subjective one

giving the percentage of words correctly recognized by the system. Along with this, two measures of phonological distances between stimuli and ASR system's outputs were considered:

- Levenshtein distance [24]. It can be used to count the minimal number of operations – in terms of symbols addition, deletion or substitution – that are needed to transform a string a in another string b. As an example, the distance between the phonemes sequences /p@ti/ (*petit*) and /apeti/ (*apétit*) equals 2, since transforming /p@ti/ into /apeti/ requires two operations: 1) the addition of phoneme /a/ and 2) the substitution of /@/ by /e/;

- Weighted Levenshtein distance (referred to as *WL distance* below). The weighting is performed depending on the phonological nature of substituted phonemes. This measure takes into account the fact that two phonemes can be more or less close, depending on the number of distinctive features that they share. In the WL distance calculation, the cost of the substitution of a consonant by another one is equal to the ratio of distinctive features they share together (in the present work the features presented in [25] were considered). For example, replacement of the phone /p/ by /b/ has a cost of 0.125 since only 1 feature over 8 distinguishes these two phonemes – namely, the voicing feature. The same calculation is carried out for the substitution of vowels. The cost of the substitution of a consonant by a vowel – and conversely – is maximum, that is equal to 1.

These two distances have been normalized as a function of the number of phonemes constituting each stimulus.

6. Results

6.1. Human scores

Table 3 shows mean subjective intelligibility scores (% recognized words) for both isolated disyllabic words and sentences stimuli. Intelligibility scores globally follow the same tendencies, although a slightly more pronounced ceiling effect can be observed in the sentence repetition task.

Table 3: Human scores

| Degradation | % recognized words (standard deviation) | |
	Isolated words	Sentences
1	97.8 (14.8)	98.9 (8.9)
2	97.8 (14.8)	97.7 (8.3)
3	92.2 (26.9)	98.6 (7.5)
4	91.1 (28.6)	96.3 (9.1)
5	84.4 (36.4)	86.8 (23.7)
6	70.0 (46.0)	69.0 (37.0)
7	32.2 (46.9)	32.1 (37.3)
8	1.1 (10.5)	1.9 (6.9)
9	0 (0)	0 (0)

6.2. Automatic scores

Figure 3 illustrates automatic word recognition scores with the LIUM-62k vs. with BIG-DET-N configurations. A better correlation with human scores is found when using BIG-DET-N instead of LIUM-62k (.89 vs. .83, respectively). For this reason phonological distances between stimuli and the ASR system outputs have been calculated using BIG-DET-N. Best correlation with human scores is found with WL distances ($r = -.95$), correlation with Levenshtein distance being slightly weaker ($r = -.94$). Also, it can be observed from figure 3 that all recognition scores drop quite rapidly when passing from degradation condition 4 to 5.
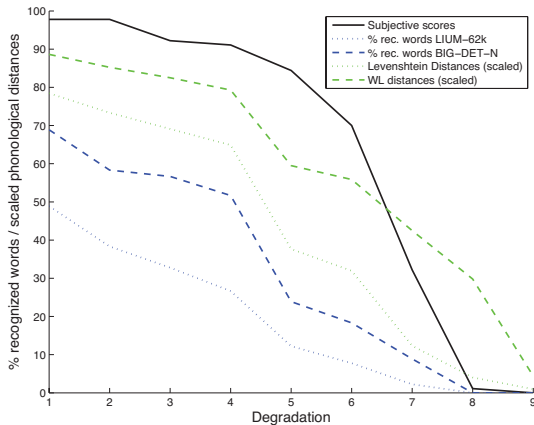
Figure 3: *Results for automatic recognition of isolated words*

Figure 4 represents automatic word recognition scores with LIUM-62k, LIUM-RES-LEX and FSG configurations. The use of a FSG leads to a greater correlation with human scores ($r = .97$) than with LIUM-RES-LEX ($r = .95$) or LIUM-62k ($r = .87$). As a consequence phonological distances between stimuli and the ASR system output have been calculated using FSG. A higher correlation with human scores is found with Levenshtein distances ($r = -.97$) than with WL Distance ($r = -.94$). As for isolated word scores, important drops can be observed at degradation 5.

7. Discussion and future work

As a main result of this study, strong correlations were observed between automatic and human speech recognition scores when dealing with stimuli simulating presbycusis at various severity grades, indicating that automatic measures designed for this task closely follow subjective scores tendencies. Nevertheless,
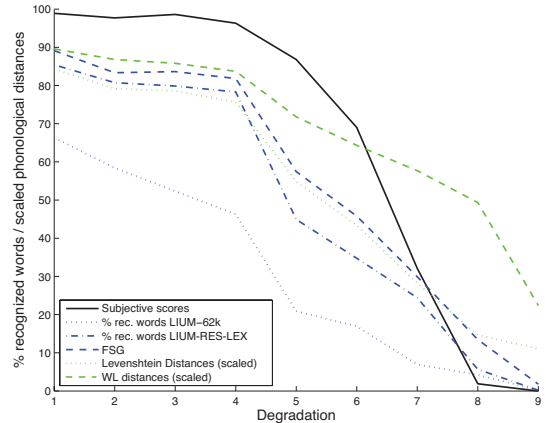
Figure 4: *Results for automatic recognition of sentences*

differences between the ASR system performances and human scores could be observed. In particular, the abrupt drops in automatic scores for degradation 5 were not found in human scores. These drops may be related to the severity steps defined in the algorithm of [12]: passing from degradation 4 to 5 is equivalent to passing from a moderate to a severe hearing loss simulation (see table 1), and implies the application of spectral smearing in a more marked way: at each severity step formant peaks become less marked, and the ASR system may have a sudden greater difficulty in processing speech stimuli. This would explain why a small drop can also be noticed in the automatic scores when passing from degradation 1 to degradation 2, which implies an increase in severity grade (from mild to moderate).

Because these drops are not observed in human scores, it may be hypothesized that listeners manage to cope with this loss of information thanks to top-down cognitive processes, i.e., making use of higher representations. In [8] a similar behavior in participants was observed: results showed that spectral smearing has no significant effect on speech intelligibility in the absence of noise; on the contrary, spectral smearing reduces significantly speech intelligibility in noise. To better reflect human scores in silent conditions, a solution could be to train the ASR system in order to make it less sensitive to spectral smearing. The observations made by [8] reinforce our aim to conduct the same kind of study with speech stimuli diffused in noise.

Along with the study of prediction of speech intelligibility in noise, future work will focus on the prediction of sentences comprehension, that is, interpretation of sentences by listeners. For collecting subjective measures a matching task between sentences from French version of HINT and images illustrating the sentences and other sentences with lexico-syntactic ambiguities is envisaged. Also, the possibility of directly measure comprehension through listeners' response to oral commands, such as in [3] is considered. On the automatic side, work will be concentrated on the recognition of keywords that help listeners to resolve syntactic or lexical ambiguities.

9. References

[1] B. C. J. Moore, *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues.* Wiley, 2007.

[2] ANSI S3.5, *Methods for the Calculation of the Speech Intelligibility Index.* American National Standards Institute, 2007.

[3] L. Fontan, J. Tardieu, P. Gaillard, V. Woisard, and R. Ruiz, "Relationship Between Speech Intelligibility and Speech Comprehension in Babble Noise," *Journal of Speech, Language and Hearing Research*, accepted.

[4] K. C. Hustad, "The Relationship Between Listener Comprehension and Intelligibility Scores for Speakers With Dysarthria," *Journal of Speech, Language and Hearing Research*, vol. 51, no. 3, pp. 562–573, 2008.

[5] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS – A System for the Automatic Evaluation of Voice and Speech Disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[6] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "Evaluation of Speech Intelligibility for Children With Cleft Lip and Palate by Means of Automatic Speech Recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.

[7] U. Ariöz, *Developing Subject-Specific Hearing Loss Simulation to Apply Different Frequency Lowering Algorithms for the Enhancement of Sensorineural Hearing Losses.* Middle-East Technical University, 2012, Ph.D. thesis accessible at http://etd.lib.metu.edu.tr/upload/12614929/index.pdf.

[8] T. Baer and B. C. J. Moore, "Effects of Spectral Smearing on the Intelligibility of Sentences in the Presence of Noise," *Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 1229–1241, 1993.

[9] B. C. J. Moore and B. R. Glasberg, "Simulation of the Effects of Loudness Recruitment and Threshold Elevation on the Intelligibility of Speech in Quiet and in Background of Speech," *Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 2050–2062, 1993.

[10] J.-E. Fournier, *Audiométrie vocale : les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités.* Maloine, 1951.

[11] V. Vaillancourt, C. Laroche, C. Mayer, C. Basque, M. Nali, A. Eriks-Brophy, S. D. Soli, and C. Giguère, "Adaptation of the HINT (Hearing in Noise Test) for Adult Canadian Francophone Populations," *International Journal of Audiology*, vol. 44, no. 6, pp. 358–369, 2005.

[12] Y. Nejime and B. C. J. Moore, "Simulation of the Effect of Threshold Elevation and Loudness Recruitment Combined With Reduced Frequency Selectivity on the Intelligibility of Speech in Noise," *Journal of the Acoustical Society of America*, vol. 102, pp. 603–615, 1997.

[13] K. J. Cruickshanks, T. L. Wiley, T. S. Tweed, B. E. Klein, R. Klein, J. A. Mares-Perlman, and D. M. Nondahl, "Prevalence of Hearing Loss in Older Adults in Beaver Dam, Wisconsin. The Epidemiology of Hearing Loss Study," *American Journal of Epidemiology*, vol. 148, no. 9, pp. 879–886, 1998.

[14] L. Fontan, C. Magnen, J. Tardieu, and P. Gaillard, "Simulation des effets de la presbyacousie sur l'intelligibilité et la compréhension de la parole dans le silence et dans le bruit," in *30ème édition des Journées d'étude sur la Parole (JEP 2014)*, Le Mans, 2014.

[15] L. Fontan, C. Magnen, J. Tardieu, I. Ferrané, J. Pinquier, J. Farinas, P. Gaillard, and X. Aumont, "Comparaison de mesures perceptives et automatiques de l'intelligibilité de la parole : cas de la parole dégradée par une simulation de la presbyacousie," *Traitement Automatique des Langues*, accepted.

[16] K. Seymore, S. Chen, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern *et al.*, "The 1997 CMU Sphinx-3 English Broadcast News Transcription System," in *Proceedings of the 1998 DARPA Speech Recognition Workshop*, 1998, pp. 55–59.

[17] Y. Estève, *Traitement automatique de la parole: contributions.* Université du Maine, 2009, mémoire d'habilitation à diriger des recherches.

[18] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news," in *Proc. Eurospeech*, Lisbon, Portugal, September 2005, pp. 1653–1656.

[19] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts," in *Proceedings of Interspeech '09*, Brighton, United Kingdom, 2009, pp. 2583–2586.

[20] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[21] S. Galliano, E. Geoffrois, G. Gravier, J. F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News," in *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy, 2006, pp. 315–320.

[22] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker Normalization on Conversational Telephone Speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, USA, 1996, pp. 339–341.

[23] B. New, M. Brysbaert, J. Veronis, and C. Pallier, "The Use of Film Subtitles to Estimate Word Frequencies," *Applied Psycholinguistics*, vol. 28, no. 4, pp. 661–677, 2007.

[24] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[25] M. Riegel, *Les sons du français : phonétique et phonologie.* PUF, 1994.

[26] X. Aumont and A. Wilhem-Jaureguiberry, *European Patent 2136359 – Method and Device for Measuring the Intelligibility of a Sound Distribution System.* Institut National de la Propriété Industrielle, 2009.