Interspeech 2018
2-6 September 2018, Hyderabad

Perceptual and Automatic Evaluations of the Intelligibility of Speech Degraded by Noise Induced Hearing Loss Simulation

Imed Laaridh[1], *Julien Tardieu*[2], *Cynthia Magnen*[2], *Pascal Gaillard*[3], *Jérôme Farinas*[1], *Julien Pinquier*[1]

[1]IRIT, Université de Toulouse, CNRS, Toulouse, France
[2]MSHS-T (USR 3414), Université de Toulouse, CNRS, France
[3]CLLE (UMR 5263) Université de Toulouse, CNRS, France

firstname.name@irit.fr[1], firstname.name@univ-tlse2.fr[2,3]

Abstract

This study aims at comparing perceptual and automatic intelligibility measures on degraded speech. It follows a previous study that designed a novel approach for the automatic prediction of Age-Related Hearing Loss (ARHL) effects on speech intelligibility. In this work, we adapted this approach to a different type of hearing disorder: the Noise Induced Hearing Loss (NIHL), i.e., hearing loss caused by noise exposure at work. Thus, we created a speech corpus made of both isolated words and short sentences pronounced by three speakers (male, female and child) and we simulated different levels of NIHL. A repetition task has been carried out with 60 participants to collect perceptual intelligibility scores. Then, an Automatic Speech Recognition (ASR) system has been designed to predict the perceptual scores of intelligibility. The perceptual evaluation showed similar effects of NIHL simulation on the male, female and child speakers. In addition, the automatic intelligibility measure, based on automatic speech recognition scores, was proven to well predict the effects of the different severity levels of NIHL. Indeed, high correlation coefficients were obtained between the automatic and perceptual intelligibility measures on both speech repetition tasks: 0.94 for isolated words task and 0.97 for sentences task.

Index Terms: Speech intelligibility metric, automatic speech recognition, hearing disorders, noise induced hearing loss simulation.

1. Introduction

More than 3 million workers in France are exposed, within their professional environment, and for extended periods, to high and potentially harmful levels of noise. Indeed, according to the French noise information centre (CidB), near 1200 cases of Noise Induced Hearing Loss (NIHL) are diagnosed each year. This hearing disorder is characterised by the presence of dead cochlear zones (i.e. deterioration of some parts of the hair cells preventing tonotopic coding). The appearance of these dead zones is accompanied by tinnitus and hyperacusis phenomena resulting in speech comprehension problems in both silence and noise conditions.

More specifically, this disorder translates into a sensorineural loss in the high frequency region as well as a localised notch around 4 kHz [1]. The severity of the loss and the auditory notch depends on the severity of the trauma causing the disorder and can evolve in time affecting other frequency bands (more precisely the conversational bands) surrounding the areas most affected initially. These alterations, when left untreated, can compromise the communication capacities of those affected and cause adverse effects on their personal and professional lives such as isolation [2] and depression [3]. One of the main solutions to hearing diseases is the use of Hearing Aids (HA). These tools can, by amplifying certain frequency bands, give back a better audibility to the affected patients. However, as reported in [4], about 40% of the patients equipped with this type of tools never (or rarely) use it. This rejection can be related, among other causes, to the lack of tuning and specific settings of the HA to each particular user.

In clinical practice, the evaluation of hearing disorders is generally performed using audiograms and perceptual tests. Generally, this evaluation consists in a transcription task of some linguistic contents (words or sentences) by the patients. The capacity (or lack thereof) of the listener to correctly recognise and transcribe the uttered sequences can then be used as a measure of his/her hearing ability. These tests can be used both for the diagnosis of the disorder and the tuning of the HA in order to obtain the best intelligibility gain for the patients.

However, several limitations can be associated with this type of perceptual evaluation. For example, the often manual implementation and result processing methods causes these tests to be very long and time consuming. Also, and in order to obtain more robust measures and suppress listener-dependant behaviour, the use of a jury of many listeners is usually recommended. Due to these characteristics, the clinical application of perceptual evaluation, especially during HA fitting phases, becomes very limited. In addition, the linguistic material usually used in these tests is often quite limited (example: the list of words of Fournier [5]) and not diversified. As a result, the patients can become very quickly familiar with the linguistic content of the tests. This, coupled with the fact that these evaluations have to be repeated several times during the HA fitting phase, can highly compromise their results [6].

The rest of this paper is organised as follows. In section 2, the context and objectives of this project are presented. Section 3 presents the methodology implemented in this work for the simulation of NIHL and the Automatic Speech Recognition (ASR) system used. Section 4 presents the different results of both the perceptual and the automatic intelligibility evaluation methods whereas section 5 provides some conclusions and directions for future work.

2. Context and objectives

This study is a part of a larger multidisciplinary project involving different experts in speech and language processing; speech therapists, computer scientist and ear, nose and throat specialists. This project is associated with previous work that proposed

10.21437/Interspeech.2018-1264

and validated a methodology for the simulation and automatic evaluation of the effects of age related hearing loss on speech intelligibility [7].

This study aims to adapt this methodology to other hearing disorders, specifically NIHL. The proposed approach will enable audiologists to automatically evaluate (on the basis of the ASR results) the NIHL effects on the intelligibility and thus to better tune and adjust of the HA proposed to the patients. On a broader level, this project also aims to adapt the proposed methodology to other languages than French (such as English) and other degradation conditions (such as speech produced in noise condition).

In another context, and considering the relation between speech production and hearing disorders, the proposed methodology could be used for the objective evaluation of the speech intelligibility for patients suffering from speech and communication disorders: dysarthria, head and neck cancers, etc. Indeed, the objective and reproducible nature of this evaluation would be very useful in patient's rehabilitation context and during longitudinal studies of such speech disorders.

3. Methodology

The proposed methodology for the simulation and evaluation of NIHL consists of 3 phases reported in figure 1.

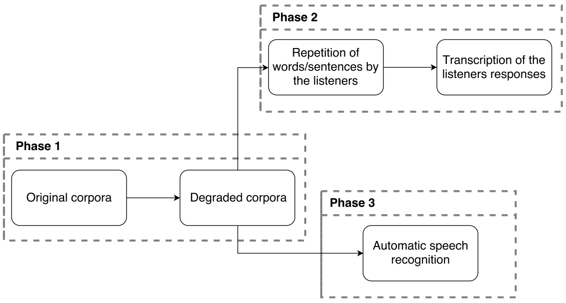
Phase 2
Repetition of words/sentences by the listeners
Transcription of the listeners responses
Phase 1
Original corpora
Degraded corpora
Phase 3
Automatic speech recognition

Figure 1: *Representation of the different phases of the proposed methodology.*

The first phase describes the corpus constitution; words from the Fournier list and sentences from the Hearing In Noise Test (HINT) were recorded by 3 speakers (man, women and child). Different levels of degradation associated with NIHL were simulated on the recordings in both silence and noise conditions (see subsection 3.1). The second phase of the methodology is the perceptual evaluation of the intelligibility of the degraded speech recordings by a jury of 61 listeners. The evaluation protocol is described in the subsection 3.2. The same recordings were then used in the third and final phase of the methodology: an automatic speech recognition task in order to compute automatic intelligibility measures. The ASR system used in this work is described in the subsection 3.3.

3.1. Phase 1: speech material

The linguistic material consists in audio recordings of two speech production tasks: isolated words (T1) and sentences (T2). The words used in T1, 60 words (6 lists of 10 words), were extracted from the word lists proposed by Fournier in 1951 for the perceptual evaluation of the speech intelligibility [5]. These lists are widely used by experts and professionals for the evaluation of patient's hearing. All the words start with a consonant and were produced with the following form: article + noun (example: ”le parfum” (*the perfume*)). The 60 sentences used in T2 (3 lists of 20 sentences), were taken from the French version of HINT [8]. The selected sentences formed single assertive clauses while maintaining a simple syntactic structure (example: ”Il vit dans la jungle” (*He lives in the jungle*)).

Three speakers were recruited for the recording session: a man (46 years old), a woman (47 years old) and a girl (12 years old). All the speakers had French as their mother tongue and produced all the sentences (60) and words (60) on the lists. The recordings were performed in an audio-metric booth (PETRA[1]) using a Sennheiser MD46 omnidirectional microphone, a TASCAM DM-3200 mixing console and a MacPro computer equipped with the Reaper software.

In order to simulate the presence of ambient noise in the recordings produced in silence condition, we used a ”babble” background noise following the method described in [7]. The babble was mixed with the recordings with a Signal to Noise Ratio (SNR) of 5 dB. Then, both speech stimuli (in silent and noise conditions) were degraded in order to simulate the NIHL effects. The degradation simulation was carried out using MATLAB following the algorithm developed initially in [9]. On the basis of real collected data[2], 10 audiograms (see figure 2) were simulated ranging from normal hearing level (level 1) to severe NIHL (level 10).

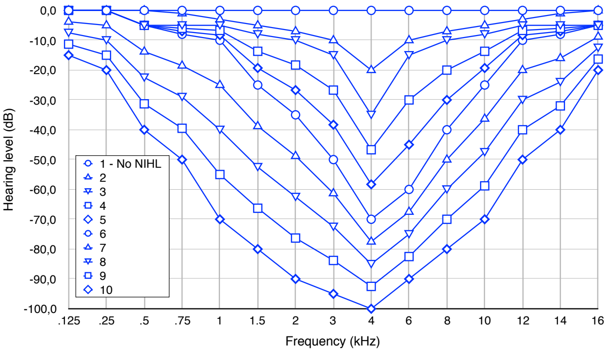

Figure 2: *Audiograms used in the simulation of 10 levels NIHL, expressed in hearing loss (dB) for each frequency band (kHz).*

3.2. Phase 2: perceptual evaluation

The perceptual evaluation of intelligibility consisted in a task of listening and repetition of the different stimuli presented to participants at the different NIHL levels in both quiet and noise.

First, an audiogram was performed for each participant to monitor his/her hearing level using the AudioConsole software. Participants with an average hearing loss superior to 15 dB at frequencies 1, 1.5, 2, 3 and 4 kHz were excluded from the experiment. Then, a training session with 10 words was run to get the participants familiar with the demanded task. Finally, each participant listened to 60 different stimuli in the 60 conditions randomly applied: 10 NIHL levels * 3 speakers * 2 noise conditions. Each participant was asked to listen carefully, and then to repeat (produce) what he/she heard (word or sentence). All the participants were seated one meter away from two Focal Solo 6 BE speakers and in front of the microphone used to collect their productions of what they heard in the audio-metric booth

[1] http://petra.univ-tlse2.fr

[2] https://www.uvmt.org/sections.php?op=printpage&artid=568

(PETRA). The sound level of the clean non-degraded stimuli in quiet (degradation level 1 in quiet) was set to 60 dBA. The responses of the participants were then transcribed by annotators (among the writers of the paper) in order to compute perceptual intelligibility measures.

Two different groups of participants were recruited for each test: T1 (words) and T2 (sentences). 31 participants were selected for T1: 20 women and 11 men (mean age=20.5 years old, $\sigma = 1.8$). 30 participants were selected for T2: 19 women and 11 men (mean age=21.2 years old, $\sigma = 2.7$). All participants had French as their mother tong, no uncorrected sight problems, and were students in fields other than music, language sciences, foreign languages and psychology.

3.3. Phase 3: automatic evaluation approach

We used an ASR system based on the Sphinx-3 [10] tool distributed by the Carnegie Mellon University (CMU). It is important to note that, unlike in classic ASR applications, this work does not aim to propose and improve the performances of the automatic speech recognition in degraded conditions (in terms of Word Error Rate - WER). The aim of this work is to use an ASR system and to propose an automatic approach able to simulate and reproduce the human perception behaviour when facing distortions associated with NIHL.

The acoustic models used in this work were proposed by the computer science laboratory of the university of Maine, France (LIUM) [11] and were trained over several hours of radio-recordings from the ESTER corpus [12]. The models used are contextual, composed of 35 phonemes and 5 types of pauses/silences (resulting in 5725 phonemes in context) where each state is represented by a Gaussian Mixture Model (GMM) of dimension 22. All the recordings are sampled at 16 kHz and PLP [13] features are used. Also, and since the acoustic speech models were trained on male speech only, the ASR system behaviour was not adapted to the speech productions of both women and child speakers. To overcome this limitation, a Vocal Tract Length Normalisation (VTLN) [14] was used. This adaptation is based on the hypothesis of an existing linear relation between the length of the vocal tract of a speaker and his/her speech formant areas.

Two models were used in the word recognition task T1: (1) a basic trigram model (called BM) trained on the ESTER2 corpora and a lexicon of around 62k words (2) a bigram model (called LM-T1) reflecting the particular syntactic composition of the used list of words (article + noun) and trained on only disyllabic words starting with a consonant (around 15k). The frequencies used for each form were those defined in [15] based on movies subtitles from the Lexicon 3.8 database. Several language model/lexicon configurations were used for the sentence recognition task T2:

- LM-T2: a trigram model trained on only a sub-part of the ESTER2 corpus (the sentences containing at least one word from the T2 sentences);
- BM: language model BM with a 62k word lexicon;
- BM-HINT: language model BM with a lexicon containing only the word from the HINT test (260 sentences);
- BM-T2: language model BM with a lexicon containing only the words from the T2 60 sentences.

4. Results

4.1. Perceptual evaluation of speech intelligibility

The perceptual evaluation of intelligibility is measured as follows. For each word of T1, the score is equal to 1 if the word is correctly repeated by the participant (only the noun is considered, the article preceding each word is not taken into account) and 0 if the word is partially/not repeated. For each sentence of T2, the score is equal the ratio between the number of fully repeated words and the total number of words in the heard sentence (in percent). For each test, the ratio of the fully repeated words is considered as the perceptual evaluation measure of intelligibility. Figure 3 depicts the perceptual intelligibility measure for each degradation level and for each test.

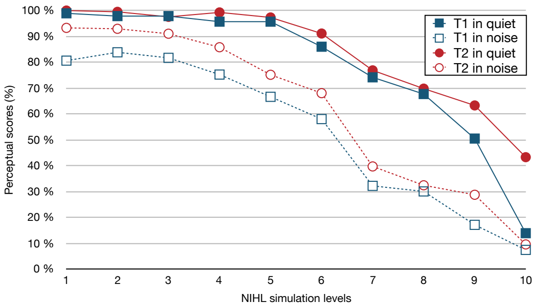

Figure 3: *Perceptual intelligibility measure per degradation level in both silence and noise conditions for words (T1) and sentences (T2).*

The results were analysed using a generalised linear mixed model for T1 and a linear mixed model for T2. In T1, the degradation level and the noise condition were found to have significant effect ($p < 0.001$) whereas the type of speaker (gender and age) was shown to be non-significant ($p = 0.63$). In T2, the degradation level and the noise condition were also found to have significant effect ($p < 0.001$), the type of speaker was not significant ($p = 0.1$). These results are consistent with those obtained earlier in a work studying the effects of simulated ARHL on words intelligibility [7].

4.2. Automatic evaluation of speech intelligibility

In this experiment, we only applied the automatic intelligibility evaluation approach to the stimuli in clean condition (no noise). This choice aims at validating the approach and its potential on clean speech before testing it facing more challenging conditions. This will also allow a better comparison of the approach behaviour on NIHL with previous results [7].

Figure 4 depicts the mean automatic and perceptual recognition rates for the 3 speakers per simulated degradation level for the word recognition task. Two automatic recognition strategies were used. The first, stricter, considers that a word is recognised only if it is the best match proposed by the ASR system. The second considers that a word is recognised if it appears on the list of the 10 best proposal of the ASR system.

Observing figure 4, we find that with the exception of the degradation level 10 (most severe), the automatic word recognition rate is lower than the perceptive rate. Indeed, the mean recognition rate is 77.8% for human listeners comparing to 60.9% and 52.4% for the ASR system using LM-T1 and BM language models respectively (first recognition strategy). The capacity of the ASR system, contrary to human perception, to compensate the most severe degradation (showed by the higher

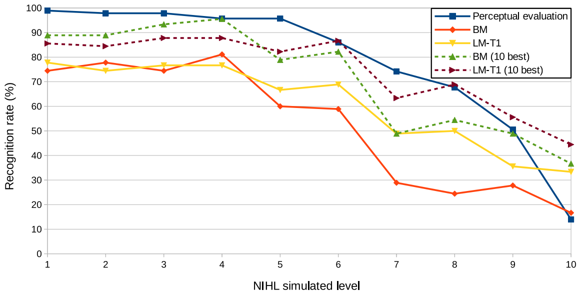

Figure 4: *Perceptual and automatic recognition rates (using both language models) per simulated NIHL degradation level for task T1 (words).*

recognition rate on the degradation level 10) requires a more in-depth investigation of the speech stimuli. Also, and as expected, we observe that the ASR configuration using the LM-T1 language model, more adapted to the structure of the word list used in our experimental protocol, reaches better recognition rates than baseline model BM.

Considering both automatic recognition strategies, the automatic approach behaviour follows the same trend of the perceptual evaluation: the higher the simulated degradation level is, the higher the WER is. Considering more precisely the evolution of these measures according to the NIHL degradation severity, we observe that the human perception is able to well compensate the distortions observed on the 5 first simulated degradation levels. Then, starting at level 5, the perceptual intelligibility is lost with an almost linear slope, with a much important dip in intelligibility observed between levels 9 (50.5%) and 10 (14%). The automatic intelligibility measure, present on the other hand, a different evolution, more sensitive to low degradation (recognition loss between levels 4 and 5) and more resistant to the severe degradation (level 10).

Figure 5 depicts the mean automatic and perceptual recognition rates for the 3 speakers per simulated degradation level using the different language model/lexicon configurations for the sentence recognition task. Once again, the perceptual intelligibility measures are higher than the automatic scores (average of 83.8%). The average automatic recognition rate varies depending on the language model/lexicon configuration reaching 50.3%, 47.3%, 56.9% and 61.6% for BM, LM-T2, BM-HINT and BM-T2 respectively. The higher recognition rates computed when using BM-HINT and BM-T2 were expected considering the limited lexicon used in these configurations. Figure 5 shows that automatic and perceptual intelligibility measures presents the same trend observed earlier on the T1 task and proves the consistency of the proposed automatic intelligibility measure. It is also interesting that, as observed earlier and contrary to human perception, the ASR system better compensate the most severe degradation.

In order to evaluate the capacity of the automatic intelligibility measure to produce similar measures to the perceptual evaluation, we computed the Pearson correlation coefficient between both measures. However, and to compensate the ceiling effect observed on the scores, we transformed both measures into rational-arcsine units (RAU) [16]. Table 1 reports the correlation computed between both transformed measures.

Considering the word recognition task, the correlation coefficient between both measures reaches 0.94 when using the LM language model. This rate confirms the capacity of the automatic approach to replicate the behaviour of the human perception when facing simulated NIHL degradations. The correlation coefficient between automatic and perceptual measures is higher than 0.96 for all the language model/lexicon configurations reaching 0.974 when using the BM model and the limited lexicon of the T2 task. These rates are also consistent with the results previously obtained when studying ARHL (0.97 and 0.98 on T1 and T2 tasks respectively) [7].

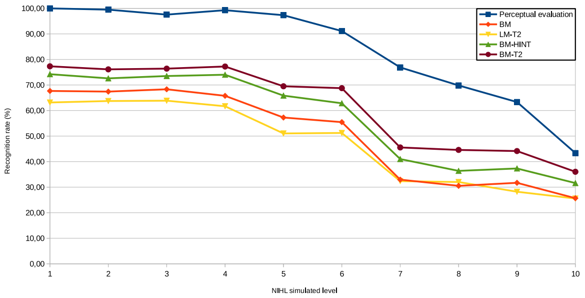

Figure 5: *Perceptual and automatic recognition rate (using the different language model/lexicon configurations) per simulated NIHL degradation level for task T2 (sentences).*

Table 1: *Pearson correlation between automatic and perceptual intelligibility measures ($p < 0.001$).*

	Model	Pearson correlation
Word task T1	BM	0.906
	LM-T1	0.943
Sentence task T2	BM	0.965
	LM-T2	0.963
	BM-HINT	0.972
	BM-T2	0.974

This proved relation between the perceptual and the automatic intelligibility measures confirms the benefit of the proposed approach for audiologists and its potential use for the tuning and specific adaptation of HA for patients suffering from NIHL.

5. Conclusions

This work aims to study the effects of NIHL simulation on speech intelligibility using both word and sentence recognition tasks. The perceptual evaluation performed has shown that these effects are independent of the gender and age of the speakers and was then used as a reference to evaluate the automatic speech intelligibility measure. The proposed approach, based on ASR, presented a similar trend and behaviour compared to the perceptual measure (correlation of 0.94 and 0.97 on the word and sentence tasks respectively). These results validate the proposed approach and confirm its potential use for the tuning of HA.

Future work will study if the approach could be generalised to other languages (English) and whether it can be applied in more difficult environments (noise conditions).

6. Acknowledgements

This study was carried out as part of the project 14054092 CLE 2015 PHONICS: Intelligent Electronic Device for Measuring Speech Comprehension , funded by the Midi-Pyrénées region. The partners are ”Maison des Sciences de l'Homme et de la Société de Toulouse” (MSHS-T), IRIT and Archean Technologies. We deeply thank Christian Füllgrabe for his help and advice for the methodology.

7. References

[1] D. McBride and S. Williams, “Audiometric notch as a sign of noise induced hearing loss,” *Occupational and Environmental Medicine*, vol. 58, no. 1, pp. 46–51, 2001.

[2] W. J. Strawbridge, M. I. Wallhagen, S. J. Shema, and G. A. Kaplan, “Negative consequences of hearing impairment in old age: a longitudinal analysis,” *The Gerontologist*, vol. 40, no. 3, pp. 320–326, 2000.

[3] B. Gopinath, J. J. Wang, J. Schneider, G. Burlutsky, J. Snowdon, C. M. McMahon, S. R. Leeder, and P. Mitchell, “Depressive symptoms in older adults with hearing impairments: the blue mountains study,” *Journal of the American Geriatrics Society*, vol. 57, no. 7, pp. 1306–1308, 2009.

[4] L. Vestergaard Knudsen, M. Öberg, C. Nielsen, G. Naylor, and S. E. Kramer, “Factors influencing help seeking, hearing aid uptake, hearing aid use and satisfaction with hearing aids: A review of the literature,” *Trends in amplification*, vol. 14, no. 3, pp. 127–154, 2010.

[5] J. E. Fournier, “Audiométrie vocale : les épreuves d’intelligibilité et leurs applications au diagnostic, à l’expertise et à la correction prothétique des surdités.” Paris, 1951.

[6] K. C. Hustad and M. A. Cahill, “Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech,” *American Journal of Speech-Language Pathology*, vol. 12, no. 2, pp. 198–208, 2003.

[7] L. Fontan, I. Ferrané, J. Farinas, J. Pinquier, J. Tardieu, C. Magnen, P. Gaillard, X. Aumont, and C. Füllgrabe, “Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss,” *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 9, pp. 2394–2405, 2017.

[8] 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: Adaptación del hint (prueba de audición en ruido) para poblaciones de adultos canadienses francófonos,” *International Journal of Audiology*, vol. 44, no. 6, pp. 358–361, 2005.

[9] Y. Nejime and B. C. Moore, “Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise.” *The Journal of the Acoustical Society of America*, 1997.

[10] 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 *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[11] P. Deléglise, Y. Esteve, S. Meignier, and T. Merlin, “The LIUM speech transcription system: a CMU Sphinx III-based system for french broadcast news,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[12] S. Galliano, G. Gravier, and L. Chaubard, “The ESTER 2 evaluation campaign for the rich transcription of french radio broadcasts,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[13] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[14] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, “Speaker normalization on conversational telephone speech,” in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 339–341.

[15] 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.

[16] G. A. Studebaker, “A rationalized arcsine transform,” *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 3, pp. 455–462, 1985.