Objective Intelligibility Assessment of Text-to-Speech Systems Through Utterance Verification

Raphael Ullmann[1,2], Ramya Rasipuram[1], Mathew Magimai.-Doss[1], and Hervé Bourlard[1,2]

[1]Idiap Research Institute, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{raphael.ullmann, ramya.rasipuram, mathew, bourlard}@idiap.ch

Abstract

Objective assessment of synthetic speech intelligibility can be a useful tool for the development of text-to-speech (TTS) systems, as it provides a reproducible and inexpensive alternative to subjective listening tests. In a recent work, it was shown that the intelligibility of synthetic speech could be assessed objectively by comparing two sequences of phoneme class conditional probabilities, corresponding to instances of synthetic and human reference speech, respectively. In this paper, we build on those findings to propose a novel approach that formulates objective intelligibility assessment as an utterance verification problem using hidden Markov models, thereby alleviating the need for human reference speech. Specifically, given each text input to the TTS system, the proposed approach automatically verifies the words in the output synthetic speech signal and estimates an intelligibility score based on word recall statistics. We evaluate the proposed approach on the 2011 Blizzard Challenge data, and show that the estimated scores and the subjective intelligibility scores are highly correlated (Pearson's $|R| = 0.94$).

Index Terms: Speech intelligibility, objective measures, text-to-speech synthesis, utterance verification, KL-divergence, KL-HMM

1. Introduction

Intelligibility is a crucial aspect in many applications that use Text-To-Speech (TTS) synthesis, for example screen readers and public address (PA) systems. More recently, synthetic speech has also been used in speech coding to achieve very low bit rates [1]. Intelligibility is most reliably assessed through subjective listening tests, but such tests are expensive and time-consuming to conduct. It is therefore desirable to have an *objective* measure that can predict subjective intelligibility scores. An objective measure also yields repeatable evaluations, making it useful for the design and optimization of TTS systems.

Most objective measures of intelligibility were designed to assess distorted human speech, e.g., due to background noise or reverberations. These measures typically compare acoustic properties of the test signal to those of the undistorted original signal [e.g., 2, 3]. The comparison of acoustic properties requires that both signals be from the same speaker. This approach is not suitable to assess the intelligibility of clean (undistorted) synthetic speech, since a more intelligible version of the recording from the same speaker may not exist.

This research was partly funded by CTI project ScoreL2 and by armasuisse, the competence center for procurement and technology within the Swiss Federal Department of Defense, Civil Protection and Sport. We gratefully acknowledge the organizers of the Blizzard Challenge for making their data available for research.

Motivated by this, an approach was proposed in [4] based on the comparison of phoneme posterior probability features of synthetic speech and human reference speech. The use of features in the *phonetic* domain makes it possible to compare speech from different speakers. Specifically, the Kullback-Leibler divergence was used to compare posterior features of human and synthetic speech recordings of the same sentences. It was shown that this method could predict significant differences in subjective intelligibility scores between TTS systems [4].

In this paper, we improve on the previous work in [4] to propose a novel approach for intelligibility assessment. We cast intelligibility assessment as an utterance verification task, where TTS test utterances are verified against an automatically generated reference phoneme posterior probability sequence. More specifically, the reference sequence is obtained through a Kullback-Leibler divergence based HMM (KL-HMM), removing the need for a reference human speech recording of the same sentence. A measure of uncertainty is estimated for each word based on the average KL-divergence between the phoneme probability sequence of the synthetic speech signal and the KL-HMM reference. The word recall for the synthetic speech utterance is then computed by thresholding the uncertainty measures. Our evaluations on the 2011 Blizzard Challenge data [5] show that the computed word recall statistics directly correlate with the word accuracy scores from subjective intelligibility tests.

This paper is organized as follows: We review existing approaches to objective intelligibility assessment in Section 2. Section 3 explains the proposed utterance verification approach, and the experimental setup is described in Section 4. We evaluate our approach in Section 5, where we demonstrate the relationship between word recall and subjective word accuracy scores. We further discuss the results and conclude in Section 6.

2. Relevant Literature

The traditional approach to objective intelligibility assessment consists in measuring specific features of the speech signal that are known to be relevant to intelligibility. For example, when speech is degraded by additive noise, it is possible to predict its intelligibility by analyzing the relative intensity of speech and noise within different frequency bands [6] and over time [7]. Objective measures that assess further signal degradations, including reverberation, speech coding or time-frequency masking, use per-band envelope intensities or spectro-temporal representations as features, and compare them to the features extracted from the original, undistorted signal [e.g., 2, 3, 8, 9]. However, in the case of synthetic speech the "degradations" are imperfections of the TTS system, and there is no "original signal" to which these acoustic features can be compared.

A possible solution is to verify whether the phone- or word-level content of the TTS signal is consistent with a reference transcription or a human speech recording of the same words. Wang et al. [10] used the decoder of an ASR system to analyze the phone graph in synthetic speech, and compared it to multiple templates of individual phones with different context. Their system requires a reference phonetic transcription for the comparison, and achieved high correlation with subjective TTS intelligibility scores. In [4], phoneme posterior probability sequences from synthetic and human speech instances of the same sentences were extracted with an Artificial Neural Network (ANN). The sequences were compared using dynamic time warping (DTW), with the Kullback-Leibler (KL) divergence of probabilities as the local score. The average DTW score was then shown to predict significant differences in subjective TTS intelligibility scores.

Finally, approaches that use Automatic Speech Recognition (ASR) in order to perform a comparison to a reference word-level transcription have been proposed to assess the intelligibility of degraded and pathological human speech [11, 12].

3. Proposed Objective Intelligibility Measure

It can be argued that listeners assess speech intelligibility based on their prior linguistic knowledge, specifically acoustic-phonetic and lexical knowledge. To avoid the influence of further factors such as sentence context, subjective tests are conducted with special test utterances, e.g., rhyming words [13] or semantically unpredictable sentences [14]. Unintelligible words can thus be seen as cases of mismatch between the observed synthetic speech signal and listener's linguistic knowledge.

In this paper, we build on this line of thought to show that this kind of mismatch can be measured through an automatic method. Specifically, we can formulate objective TTS intelligibility assessment as an *utterance verification* problem, since the utterance that the TTS system should produce is known a priori. An automatic measure of how many words can be recalled from the TTS signal can then be linked to subjective intelligibility.

Our proposed approach estimates the recall for each word by comparing the sequence of phoneme posterior probabilities in the TTS signal to a reference. Specifically, we use a Kullback-Leibler divergence-based Hidden Markov Model (KL-HMM) [15, 16] to generate the reference sequence of phoneme posterior probabilities for a given TTS input text. The two sequences are aligned and the mismatch between the TTS signal and the linguistic knowledge modeled by the KL-HMM is evaluated for each word.

The proposed approach improves on the previous work in [4], where a human speech recording of the TTS input text was used as a reference. The KL-HMM alleviates the need for human speech recordings of each tested TTS utterance, and provides the word-level segmentation. Moreover, the KL-HMM can be trained on data from multiple recordings and speakers to provide a more general reference than a single human speech recording.

The architecture of the proposed objective measure is depicted in Figure 1 and consists of the following parts:

Synthetic speech A TTS system takes as input a sequence of words $W = \{w_1, \ldots, w_m, \ldots, w_M\}$ and converts them to speech.

Spectral feature extraction Given a synthetic speech utterance, a sequence of acoustic feature observations $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N\}$ is estimated where N denotes the number of frames in the TTS speech signal. The features can be e.g., cepstral coefficients of a short-time mel-frequency or auditory spectrum (MFCC, PLP).

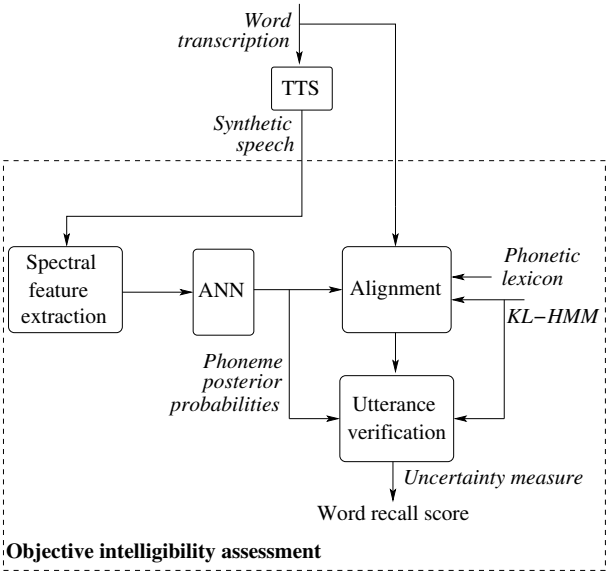

Figure 1: Architecture of the proposed objective TTS intelligibility assessment system

Posterior feature extraction The acoustic feature observation sequence X is converted into a sequence of phoneme posterior probabilities $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_n, \ldots, \mathbf{z}_N\}$ using an Artificial Neural Network (ANN) where

$$\begin{aligned}\mathbf{z}_n &= [z_n^1, \ldots, z_n^k, \ldots, z_n^K]^\top \\ &= [P(c_1|\mathbf{x}_n), \ldots, P(c_k|\mathbf{x}_n), \ldots, P(c_K|\mathbf{x}_n)]^\top,\end{aligned} \tag{1}$$

with c_k the k^{th} phoneme class of K phonemes.

Alignment In a KL-HMM, each HMM state i is parameterized by a categorical distribution $\mathbf{y}_i = [y_i^1, \ldots, y_i^k, \ldots, y_i^K]^\top$. The phoneme posterior probabilities estimated by an ANN are directly used as feature observations to train the KL-HMM.

Word level and subword state level time alignments are obtained for the input sequence of phoneme posterior probabilities Z and its word level transcription W through Viterbi alignment using a trained KL-HMM system and a phonetic lexicon. The local score $\mathrm{KL}(\mathbf{y}_i, \mathbf{z}_n)$ at each HMM state i is the KL-divergence between the state distribution $\mathbf{y}_i$ and the posterior feature $\mathbf{z}_n$, i.e.,

$$\mathrm{KL}(\mathbf{y}_i, \mathbf{z}_n) = \sum_{k=1}^{K} z_n^k \log\left(\frac{z_n^k}{y_i^k}\right) \tag{2}$$

Utterance verification An uncertainty measure $C(w_m)$ for each word w_m is computed based on the average KL-divergence between the sequence of phoneme posteriors of the synthetic speech signal and the KL-HMM subword states,

$$C(w_m) = \frac{1}{R_m} \sum_{r=1}^{R_m} \frac{1}{e_{rm} - b_{rm}} \sum_{n=b_{rm}}^{e_{rm}} \mathrm{KL}(\mathbf{y}_{s_{rm}}, \mathbf{z}_n) \tag{3}$$

with s_{rm} the r^{th} subword state in word w_m, b_{rm} and e_{rm} the begin and end indices of the frames aligned with subword state s_{rm}, and R_m the number of subword states for word w_m, respectively. The uncertainty measure $C(w_m)$ takes into account the number of frames in each subword state and the number of subword states in each word, similar to the double normalization approach for hybrid HMM/ANN systems [17].

Finally, we calculate word recall by comparing the uncertainty measure $C(w_m)$ of each word to a decision threshold τ. The value τ may be chosen such that the calculated word recall correlates best with intelligibility scores, e.g., using a small development set of subjectively scored TTS speech recordings.

Alternatively, we can follow the utterance verification formulation to select the threshold τ without subjectively scored data. Specifically, we take a hypothesis testing approach where we wish to select the threshold τ that best separates two distributions H_0 and H_1 of uncertainty scores. The H_0 hypothesis means that the target word is present in the signal, whereas H_1 means that the TTS system synthesized a signal that does not agree with our lexical and phonetic knowledge for the word.

We obtain uncertainty scores for the H_0 distribution from speech signals with known high intelligibility, e.g., undistorted human speech. Uncertainty scores for the H_1 distribution can be obtained by distorting the signal, or — more simply — by using wrong transcriptions for utterance verification. The wrong transcription can be a word that rhymes with the true word or a completely different word, depending on the intelligibility test type (i.e., rhyme test or sentence test) that we wish to model.

We compare both approaches to selecting τ in Section 5.1.

4. Experimental Setup

We evaluate the proposed objective intelligibility assessment method on the 2011 Blizzard Challenge data [5]. The data comprises speech recordings from 12 different text-to-speech (TTS) systems, referred to as systems "B" to "M". We use a set of 26 semantically unpredictable sentences (SUS) in English, for which subjective intelligibility scores are available as word error rates (WER). Each sentence contains 6–8 words. For the following evaluations, we convert WER scores to word accuracy (WA), defined as $\text{WA} = 1 - \text{WER}$. The data also includes human speech recordings of the sentences spoken by a professional speaker, referred to as system "A".

The extraction of phoneme posterior probabilities from the TTS recordings is performed with the same single hidden layer multilayer perceptron (MLP) used in [4]. The MLP is trained on 232 hours of conversational telephone speech to classify 44 English phonemes and silence, i.e., $K = 45$ output units. Training with conversational speech brings in acoustic variability that can help make the MLP more robust. The MLP inputs are 39-dimensional perceptual linear predictive cepstral features [18] with a nine frame temporal context (i.e., four frames preceding and four frames following). The MLP was trained with the QuickNet toolkit[1] by minimizing the frame-level cross entropy.

The KL-HMM system is trained on the "system A" human speech recordings and models crossword context-dependent phonemes. Each crossword context-dependent phoneme is modeled as a 3-state HMM. The phonetic lexicon required for the KL-HMM system training is obtained from the CMU pronunciation dictionary[2].

[1] http://www.icsi.berkeley.edu/Speech/qn.html

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

5. Evaluation and Results

5.1. Threshold Selection

For each TTS system, we compute uncertainty scores for the words in the SUS recordings, using the steps described in Section 3. The word recall per sentence is obtained by comparing these scores to a decision threshold τ. Our notion is that the word recall is directly related to listener's word accuracy (WA). Since recall describes the verification of expected words and WA describes the recognition of unknown words by listeners, we expect recall to be in a higher numerical range than WA.

5.1.1. Selection with a Development Set

The relationship between objective recall and subjective word accuracy (WA) is shown in Figure 2. Lines show the linear fit between recall and WA for different values of τ. We use the first 5 of the 26 SUS recordings generated with each TTS system as development set. At low threshold values (bright lines in Figure 2), the word recall is unrealistically low, and the correlation to subjective word accuracy is also poor. As the threshold is increased to yield recall values in more realistic, higher ranges, the correlation improves too (dark lines in Figure 2). The maximum correlation on the development set is obtained at $\tau = 1.22$, with recall values between 92 and 100%.

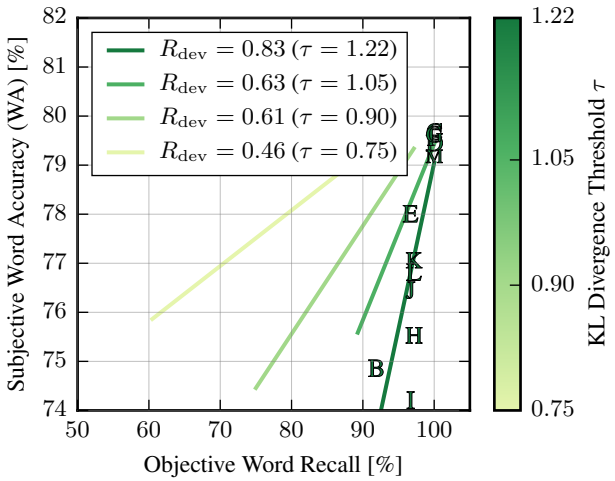

Figure 2: Relationship between word recall and word accuracy for different uncertainty thresholds τ on a development set of 5 SUS recordings per TTS system. Lines indicate the linear fit at different values of τ (for clarity, data points for the TTS systems "B" to "M" are shown for one threshold value only).

5.1.2. Selection Through a Hypothesis Testing Approach

Figure 3 shows the selection of τ from two distributions. We compute the H_0 distribution (i.e., correct word) of uncertainty scores using all human speech recordings (system "A"), which were pronounced by a professional speaker and are highly intelligible. Scores for the H_1 distribution (i.e., different word) are obtained from the same recordings of human speech, by substituting the words in the corresponding transcriptions with different words. We select the threshold value that separates H_0 from H_1 at the intersection of the fitted Beta distributions for each hypothesis. The resulting threshold $\tau = 1.05$ is close to the value obtained in the previous development set approach, but required no subjectively scored data.

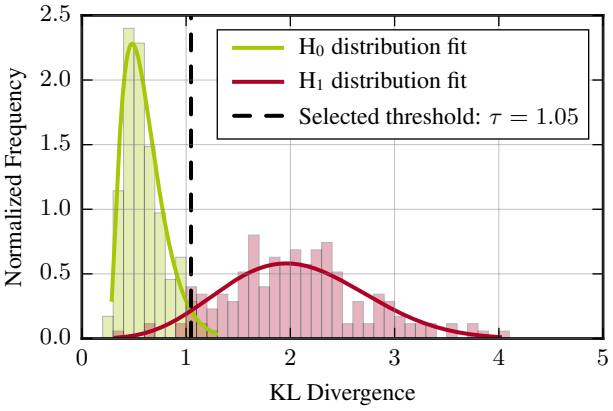

Figure 3: Selection of the decision threshold through a hypothesis testing approach. We use the objective intelligibility measure to compute word uncertainty scores for recordings of human speech (system "A") with correct and incorrect transcriptions, yielding scores for the H_0 and H_1 distribution, respectively.

5.2. Results

Table 1 shows the prediction performance of our measure ("Word Recall") for the two threshold values derived in Section 5.1. The prediction performance of the approach in [4] ("Average Distance") is also shown for comparison. The proposed measure follows a highly linear relationship to subjective intelligibility scores, as measured by the Pearson correlation coefficient R.

We evaluate the accuracy of the objective measure with the root-mean-square prediction error rmse, defined as

$$\mathrm{rmse} = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}(s_i - o_i')^2} \qquad (4)$$

with s_i and o_i' the subjective and linearly mapped objective score for TTS system i, and T the number of TTS systems, respectively. Both the correlation and accuracy metrics only degrade slightly when the threshold τ is selected through the hypothesis testing approach, indicating that the threshold could indeed be decided using data without subjective ratings.

Figure 4 depicts the relationship between subjective intelligibility scores and word recall for the first row in Table 1. Each data point "B" to "M" represents the average of 26 SUS recordings for the corresponding TTS system. The predicted intelligibility of TTS systems deviates very little from the ideal linear fit, but we see that our measure results in an outlier for human speech (system "A"). We speculate that the natural prosody of the professional speaker may have helped listeners understand some sentences (e.g., questions), resulting in an offset to synthetic speech that our measure does not account for.

Measure	Correlation (R)	Error (rmse)
Word Recall ($\tau = 1.22$)	0.94	0.69
Word Recall ($\tau = 1.05$)	0.90	0.89
Average Distance [4]	0.90	0.88

Table 1: Prediction performance of the proposed measure of synthetic speech intelligibility, evaluated by the correlation coefficient (R) and root-mean-square prediction error (rmse).

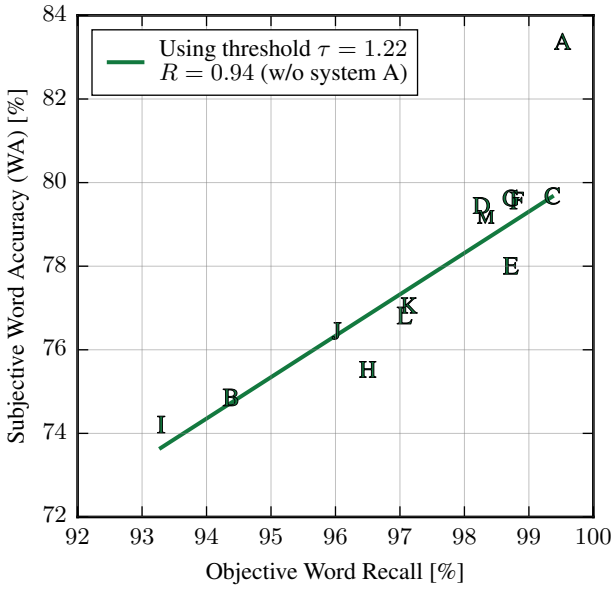

Figure 4: Prediction of subjective intelligibility from word recall. Letters indicate average scores per TTS system (system "A" is human speech). R denotes Pearson's correlation coefficient.

Conversely, we can also see in Figure 4 that subjects had slightly more difficulty understanding the output of systems H, L and E than predicted by our measure. Informal listening suggests that this may be due to irregular fluctuations in pitch and/or duration in recordings from these systems. Since our measure only uses phoneme posterior probabilities for prediction, it currently does not consider the influence of these other factors.

6. Discussion and Conclusion

We have proposed a novel formulation of objective TTS speech intelligibility assessment as an utterance verification problem using hidden Markov models. Our study shows that this method can predict subjective TTS intelligibility scores with high accuracy, while being considerably simpler than the use of a full-fledged automatic speech recognition (ASR) system. The utterance verification formulation is consistent with the way subjective intelligibility tests are conducted, in that it evaluates the mismatch between the synthetic speech signal and listener's modeled acoustic-phonetic and lexical knowledge. While subjective intelligibility tests require specially designed test utterances (e.g., nonsense words or semantically unpredictable sentences) to avoid biases due to sentence context, our objective utterance verification approach has no such requirement.

On the 2011 Blizzard Challenge data, our evaluation shows that the calculated word recall highly correlates with subjective intelligibility scores. In the objective evaluation of TTS systems, the interest lies not only in the accurate prediction of subjective scores, but also in assessing significant differences between systems. Our future work will focus on this aspect, including data from other Blizzard Challenge editions. This additional data as well as data from other domains could also be used to further train the KL-HMM system. Finally, the proposed approach could be expanded to assess other types of speech degradations, e.g., human speech distorted by low bit-rate codecs. These research directions will also be part of our future work.

7. References

[1] K.-S. Lee and R. V. Cox, "A Very Low Bit Rate Speech Coder Based on a Recognition/Synthesis Paradigm," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 482–491, 2001.

[2] J. G. Beerends, R. A. van Buuren, J. van Vugt, and J. A. Verhave, "Objective Speech Intelligibility Measurement on the Basis of Natural Speech in Combination with Perceptual Modeling," *J. Audio Eng. Soc.*, vol. 57, no. 5, pp. 299–308, 2009.

[3] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[4] R. Ullmann, M. Magimai.-Doss, and H. Bourlard, "Objective Speech Intelligibility Assessment through Comparison of Phoneme Class Conditional Probability Sequences," in *Proc. ICASSP*, Brisbane, Australia, 2015.

[5] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Proc. Blizzard Chall. Work.*, 2011.

[6] J. B. Allen, *Articulation and Intelligibility*. Morgan & Claypool, 2005.

[7] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[8] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, no. 2-3, pp. 331–348, 2003.

[9] S. Voran, "Using articulation index band correlations to objectively estimate speech intelligibility consistent with the modified rhyme test," in *IEEE Work. Appl. Signal Process. to Audio Acoust.*, New Paltz NY, USA, 2013, pp. 1–4.

[10] L. Wang, L. Wang, Y. Teng, Z. Geng, and F. K. Soong, "Objective Intelligibility Assessment of Text-to-Speech System using Template Constrained Generalized Posterior Probability," in *Proc. Interspeech*, Portland OR, USA, 2012, pp. 627–630.

[11] S. Nguyen, C. Okino, and M. Cheng, "Intelligibility and Space-based Voice with Relaxed Delay Constraints," in *IEEE Aerosp. Conf.*, 2008.

[12] A. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke, "Automatic Scoring of the Intelligibility in Patients with Cancer of the Oral Cavity," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1206–1209.

[13] A. House, C. Williams, M. Hecker, and K. Kryter, "Psychoacoustic Speech Tests: A Modified Rhyme Test," Air Force Systems Command, United States Air Force, Bedford, MA, USA, Tech. Rep., 1963.

[14] C. Benoît, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Commun.*, vol. 18, no. 4, pp. 381–392, 1996.

[15] G. Aradilla, J. Vepa, and H. Bourlard, "An Acoustic Model Based on Kullback-Leibler Divergence for Posterior Features," in *Proc. ICASSP*, Honolulu HI, USA, 2007, pp. 657–660.

[16] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task ," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 928–931.

[17] G. Bernardis and H. Bourlard, "Improving Posterior Based Confidence Measures in Hybrid HMM/ANN Speech Recognition Systems," in *Proc. ICSLP*, Sydney, Australia, 1998.

[18] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.