OBJECTIVE SPEECH INTELLIGIBILITY ASSESSMENT THROUGH COMPARISON OF PHONEME CLASS CONDITIONAL PROBABILITY SEQUENCES

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

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

ABSTRACT

Assessment of speech intelligibility is important for the development of speech systems, such as telephony systems and text-to-speech (TTS) systems. Existing approaches to the automatic assessment of intelligibility in telephony typically compare a reference speech signal to a degraded copy, which requires that both signals be from the same speaker. In this paper, we propose a novel approach that does not have such a requirement, making it possible to also evaluate TTS systems and recent very low bit rate codecs that may modify speaker characteristics. More specifically, our approach is based on comparing sequences of phoneme class conditional probabilities. We show the potential of our approach on low bit rate telephony conditions, and compare it against subjective TTS intelligibility scores from the 2011 Blizzard Challenge.

Index Terms— Speech intelligibility, Objective intelligibility, Phonemes, Artificial neural networks, KL-divergence

1. INTRODUCTION

Speech intelligibility is one of the key requirements for effective speech communication. This is not only important for human-to-human communication, but also for speech processing or transmission systems such as telephony and text-to-speech synthesis (TTS) systems to be useful. Thus, assessing the intelligibility of the output speech signal is a crucial step in the development of these systems.

Intelligibility is typically assessed through subjective listening tests, which are costly and time-consuming to conduct. It is thus desirable to develop approaches that assess speech intelligibility in an *objective* manner. Different approaches based on sample-by-sample, spectral or spectro-temporal analysis have been proposed to objectively assess the intelligibility of telephone speech. Examples are the Speech Transmission Index (STI) [1, 2], and extensions to the PESQ measure [3, 4]. These approaches however are not always extensible to assess speech synthesizers or very low bit rate (VLBR) speech codecs, which work on TTS principles (see e.g. [5]). The primary reason is that these approaches usually compare the original or reference signal to a test signal that is a degraded or distorted version of the reference signal. In telephone speech, this degradation or distortion is introduced by the codec and the transmission channel. In TTS or VLBR coding however, the reference is natural speech produced by a human speaker, and the test signal is the output of a TTS system or VLBR codec, with differences (e.g. in speaker or emotional state) that result from the speech production or synthesis mechanism[1].

The approach proposed in this work builds on recent results in template-based automatic speech recognition (ASR). In this type of ASR, a speech utterance is recognized by matching it to one of several example recordings or *templates* of possible target utterances. Soldo et al. [6] recently studied the use of synthetic speech templates (i.e., generated with a TTS system), with phoneme posterior probabilities as feature. It was observed that the approach can yield recognition performance comparable to the case where the templates are obtained using natural speech. It was also found that the performance of the system correlated with the quality of the TTS voices, more specifically with their subjective intelligibility scores.

Motivated from these observations, the present paper investigates an approach where speech intelligibility is objectively assessed by comparison of phoneme class conditional probability sequences. We demonstrate the effectiveness of the approach for the assessment of low bit rate speech codecs and TTS systems.

This paper is structured as follows: Section 2 provides a brief literature survey. The proposed approach is explained in Section 3. We present our experiments and implementation details in Section 4, and results in Section 5. Finally, we discuss and conclude in Section 6.

2. RELEVANT LITERATURE

Approaches to the objective assessment of speech intelligibility traditionally measure signal properties that were found to be important for intelligible speech, for example envelope modulations or signal-to-noise ratios within different frequency bands. A prominent example is the Speech Transmission Index (STI) [1, 2]. Calculation of the STI consists in passing a modulated noise signal through the channel under test (i.e., a given acoustic environment or speech processing system) and measuring changes to the envelope spectra. STI has been used to predict the impact of noises, reverberation, bandpass filtering and waveform coding on intelligibility.

The STI approach may not be appropriate for evaluating modern low bit rate speech codecs, which are based on a source-filter model of speech production and will process the modulated noise signal differently from actual speech. Beerends et al. [4] proposed a modified version of the PESQ model [3], which measures audible differences in the spectral domain between a reference speech signal and a degraded copy, as a new basis for objective intelligibility prediction. Comparisons between auditory spectro-temporal representations of degraded and reference speech were also proposed by Elhilali et al. [7] and Hines and Harte [8] to assess the intelligibility impact of

This research was partly funded by armasuisse, the competence center for procurement and technology within the Swiss Federal Department of Defense, Civil Protection and Sport, and CTI project ScoreL2. We acknowledge the organizers of the Blizzard Challenge for making their data available for research, and thank SwissQual AG for providing POLQA scores of Section 5.1.

[1]In VLBR coding there is an additional effect of the transmission channel.

additive noise, reverberations and phase distortions, and simulated hearing loss, respectively.

More recently, approaches have been proposed that go beyond signal or spectral level and assess intelligibility objectively at phone or phoneme level. For instance, Teng et al. [9] compared occurrences of phone bigrams (determined with an ASR system) in degraded and reference speech to assess the impact of low bit rate codecs and bit error conditions on intelligibility. By contrast, Middag et al. [10] estimated phone-level confidence scores by aggregating phone *posterior probabilities* (i.e., the probability that a target phone was pronounced) over hypothesized phone segments to perform an automated evaluation of pathological speech.

3. PROPOSED APPROACH

In the present work, we expand on ASR-based approaches to the objective assessment of speech intelligibility. More specifically, motivated from [6], the proposed approach assesses intelligibility by comparing phoneme posterior probability sequences. Given a reference speech signal and a test speech signal, the approach performs the steps outlined in Figure 1, which consist in:

1. Extraction of the reference acoustic feature sequence $A = \{\mathbf{a}^1, \ldots, \mathbf{a}^i, \ldots, \mathbf{a}^I\}$ and test acoustic feature sequence $B = \{\mathbf{b}^1, \ldots, \mathbf{b}^j, \ldots, \mathbf{b}^J\}$, where $I \lessgtr J$

2. Estimation of the reference phoneme posterior probability sequence $Y = \{\mathbf{y}^1, \ldots, \mathbf{y}^i, \ldots, \mathbf{y}^I\}$ and test phoneme posterior probability sequence $Z = \{\mathbf{z}^1, \ldots, \mathbf{z}^j, \ldots, \mathbf{z}^J\}$, where

$$\mathbf{y}^i = \left[P\left(c_1|\mathbf{a}^i\right), \ldots, P\left(c_K|\mathbf{a}^i\right)\right]^\top = \left[y_1^i, \ldots, y_K^i\right]^\top,$$
$$\mathbf{z}^j = \left[P\left(c_1|\mathbf{b}^j\right), \ldots, P\left(c_K|\mathbf{b}^j\right)\right]^\top = \left[z_1^j, \ldots, z_K^j\right]^\top,$$

   with $\sum_k y_k^i = \sum_k z_k^j = 1$, and c_k the k^{th} phoneme class out of $k \in \{1, 2, \ldots, K\}$ phoneme classes.

3. Comparison of sequences Y and Z to calculate a distance score. As noted in Step 1, the reference and test sequences may be of same or different lengths. Hence we apply Dynamic Time Warping (DTW) [11], where the local distance is the symmetric Kullback-Leibler (SKL) divergence[2],

$$\text{SKL}\left(\mathbf{y}^i, \mathbf{z}^j\right) = \frac{1}{2}\sum_{k=1}^{K} y_k^i \log_2 \frac{y_k^i}{z_k^j} + \frac{1}{2}\sum_{k=1}^{K} z_k^j \log_2 \frac{z_k^j}{y_k^i}$$

   to compute the distance between sequences Y and Z. The resulting accumulated distance, referred to as *DTW distance*, is used for intelligibility assessment.

The approach is the same for the assessment of speech codecs and TTS systems. In the case of speech codecs, the reference signal is the input signal to the codec and the test signal is the output of the codec. For the assessment of TTS systems, the reference signal is natural speech and the test signal is the TTS system output for the text corresponding to the natural speech.

[2] As demonstrated in [12], there are a number of other measures that could be used to compare probability distributions in the proposed approach.

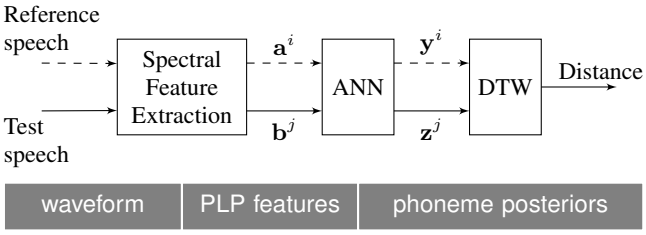

Fig. 1. Diagram of the proposed objective intelligibility measure. Phoneme posterior probabilities are estimated with an artificial neural network (ANN). The type of signal feature in each stage of the proposed approach is highlighted in gray at the bottom.

4. EXPERIMENTS

4.1. Low bit-rate coding and frame error conditions

As a first experiment, we verify how our proposed approach reacts to some signal distortions that are typical in speech telecommunications, i.e., low bit-rate coding and/or frame loss conditions. Even without subjective intelligibility scores, we can expect a trend in which lower bit rates of the same codec and increasing frame error rates both result in lower intelligibility. We test this assumption with the following conditions:

- AMR cellular telecommunication codec [13], running at eight different constant bit rates (4.75–12.2 kbps),
- EVRC-B cellular telecommunication codec [14] at the codec's standard average bit rates (4.8–9.6 kbps),
- MELP US DoD codec [15] in simple, double and triple cascaded setups (2.4 kbps),
- codec2 free open-source codec[3] operating at 2.4 kbps, with bit error rates of 0, 0.2, 0.5, 1 and 5%, and
- simulated frame loss (5, 10, 20 and 40%), by silencing randomly selected 20 ms segments of the speech signal.

We apply each condition to 12 recordings of English sentences from 12 speakers (6 male) provided in ITU-T Rec. P.501 [16]. Each recording is 2–3 seconds long, and was pre-filtered with the IRSsend telephone bandpass [17] prior to processing.

4.2. Intelligibility of synthetic speech

In a second experiment, we evaluate our model on the 2011 Blizzard Challenge data [18], which comprises speech recordings synthesized with 12 different text-to-speech (TTS) systems, referred to in the following as systems "B" to "M". Specifically, we use a subset of 26 semantically unpredictable sentences [19] in English, for which subjective intelligibility scores are provided in the form of word error rates (WER). The length of synthesized sentence recordings varies between 1 and 3 seconds, depending on the TTS system.

We chose the 2011 edition of the Blizzard Challenge, because it also included *natural* speech recordings of the sentences, pronounced by a professional voice talent. More details about the types of TTS systems, sentence material and collection of subjective scores can be found in [18].

[3] http://rowetel.com/codec2.html

4.3. Implementation

We use the same single hidden layer multilayer perceptron (MLP) used in the studies in [6, 12], trained on 232 hours of conversational telephone speech to classify 44 English phonemes and silence, i.e. 45 output units, to extract phoneme posterior probabilities ($\mathbf{y}^i$ and $\mathbf{z}^j$). The inputs to the MLP are 39-dimensional perceptual linear predictive (PLP, [20]) cepstral features[4] ($\mathbf{a}^i$ and $\mathbf{b}^j$) with four frames preceding and four frames following context, i.e., 9×39 input units. The MLP was trained with the QuickNet toolkit[5] by minimizing frame-level cross entropy. The frame size is 25 ms with a frame shift of 10 ms. In both experiments, the features are computed on telephone bandwidth.

We use the DTW implementation developed for the studies reported in [21]. In this implementation, as done in [6], the slope constraints in the DTW distance computation are:

$$D(i,j) = \mathrm{SKL}\left(\mathbf{y}^i, \mathbf{z}^j\right) + \min\left[D(i,j-1), D(i-1,j-1), D(i-2,j-1)\right]$$

where $D(i,j)$ is the accumulated distance at reference time frame i and test time frame j. However, no global constraints are applied. The final DTW distance, used for assessing intelligibility, is $D(I,J)$ normalized by the path length.

The underlying hypothesis in our experiments is that a lower overall DTW distance corresponds to higher speech intelligibility.

5. RESULTS

5.1. Low bit-rate coding and frame error conditions

We calculate average DTW distances between the original and processed recordings listed in Section 4.1, sampled at 8 kHz. Additionally, we show objective speech *quality* scores, computed with ITU-T Recommendation P.863 "POLQA" [22], the technological update to ITU-T Rec. P.862 "PESQ".

Comparing objective intelligibility and quality scores is interesting, because degradations in speech quality need not translate to lower intelligibility (e.g., robotic-sounding speech may have low perceptual quality but high intelligibility), but inversely, good intelligibility is necessary for good speech quality [23]. We should thus expect to see a range of different quality values at high predicted intelligibility, but only low quality scores when the predicted intelligibility is low.

Figure 2 compares both objective measures, with per-file scores averaged across the 12 speakers. Both the AMR and EVRC-B codecs, which operate at comparatively high bit rates, show a range of different quality values as a function of bit rate, but little variation in average DTW distance (i.e., high predicted intelligibility). The MELP codec at 2.4 kbps (single encoding pass, bright circle in Figure 2) reaches a lower quality value, but a predicted intelligibility similar to that of the two cellular telecommunication codecs. It seems indeed plausible that a codec used for military communication would be designed to maximize intelligibility. On the other hand, conditions with high DTW distance (low predicted intelligibility) are only found at low objective speech quality scores, as expected. Variations in the number of MELP encoding passes, codec2 bit errors or frame loss rates all show the expected trend. Informal listening shows that speech remains partly intelligible at 40% frame loss, but not in the codec2 condition with maximum bit error rate.

[4] $c_0 - c_{12} + \Delta + \Delta\Delta$

[5] http://www1.icsi.berkeley.edu/Speech/qn.html

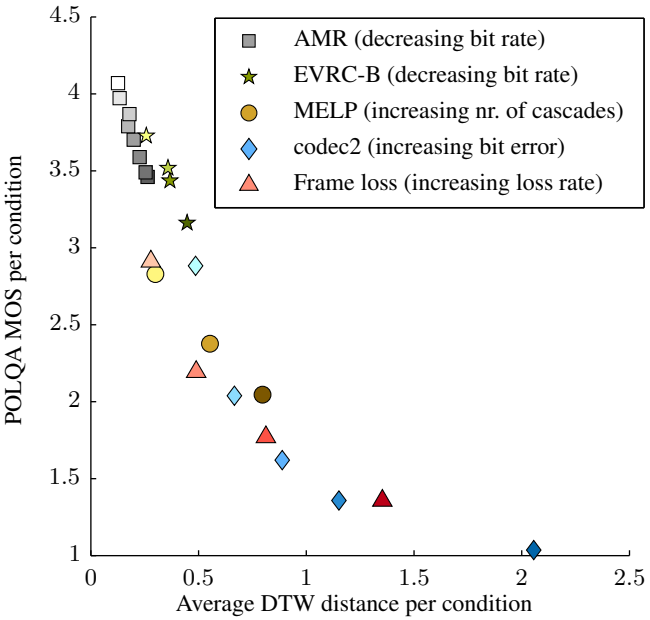

Fig. 2. Objective scores for speech *intelligibility* (proposed approach) and *quality* (POLQA), for conditions of Section 4.1. Darker data point shading corresponds to the trend given in legend parentheses. See Section 4.1 for specific codec and frame loss settings.

5.2. Intelligibility of synthetic speech

We compute the average DTW distance between synthetic and natural (human) speech recordings of 26 semantically unpredictable sentences. Table 1 compares the average distance per TTS system (*objective score*) against the average word error rate (WER) of listeners reported in [18] (*subjective score*).

The 6 most and 6 least intelligible systems are identical in both lists of Table 1, although the ordering of individual systems is not the same. However, it was found in [18] that differences in average subjective scores between two TTS systems were not always statistically significant. Evaluating the proposed approach in terms of correlation to subjective scores would not distinguish between significant and insignificant differences. Instead, we determine significant differences in *objective* scores between two TTS systems through the same statistical significance test[6] that was used in [18]. The results from both tests are overlaid in Table 2.

We see in Table 2 that all significant subjective differences are also significant with the proposed approach, and in Table 1 that the predicted rank-order of significant differences is correct. The only exception is system "J", which resulted in a non-contiguous group of subjectively equivalent systems (broken light orange bar in Table 1), indicating a possible inconsistency in subjective scores.

Moreover, objective scores for systems "B" and "H" are significantly lower than for all other systems, whereas subjects made no significant distinction between the 6 least intelligible systems (dark blue and dark red bar in Table 1, respectively). Since the proposed approach works at *phoneme* level, we hypothesize that it may be sensitive to minute differences that are not accounted for in word-level subjective evaluations.

[6] Bonferroni-corrected, paired Wilcoxon signed rank test at $p < 0.01$.

Subjective Scores		Objective Scores	
WER [%]	TTS system	Avg. distance	TTS system
16.62	natural voice	(natural used as reference)	
20.32	C	0.677	F
20.37	G	0.686	D
20.43	F	0.704	G
20.55	D	0.719	E
20.82	M	0.734	M
21.99	E	0.744	C
22.94	K	0.752	K
23.18	L	0.754	L
23.55	J	0.813	J
24.47	H	0.991	I
25.14	B	1.414	H
25.79	I	1.498	B

Table 1. Subjective and objective intelligibility scores for the 2011 Blizzard Challenge data [18] (semantically unpredictable sentences), ordered from most to least intelligible TTS system. Shaded bars indicate groups of systems with no significant differences in intelligibility scores between them.

	natural	B	C	D	E	F	G	H	I	J	K	L	M
natural		■		■	■	■	■	■	■	■	■	■	■
B	■		◘	◘	○	◘	◘		○	○	○	○	◘
C		◘						◘	◘	■			
D	■	◘						○	◘				
E	■	○						○	◘				
F	■	◘						○	◘				
G	■	◘						○	◘	■			
H	■		◘	○	○	○	○		○	○	○	○	○
I	■	○	◘	◘	◘	◘	◘	○					◘
J	■	○	■				■	○					
K	■	○						○					
L	■	○						○					
M	■	◘						○	◘				

Table 2. Significant differences in intelligibility scores between pairs of TTS systems. ■ indicates a significant difference in *subjective* scores; ○ indicates a significant difference in *objective* scores (◘ means that both subjective and objective scores are significantly different). Data for subjective scores reproduced from [18], with kind permission by the authors.

6. DISCUSSION AND CONCLUSION

We have proposed a novel approach to objective speech intelligibility assessment based on comparison of phoneme posterior probability sequences. Our investigations show that the proposed approach yields realistic results for low bit rate codec distortions, and that it is able to assess speech intelligibility for TTS systems. This second result is interesting, since a single human reference recording and acoustic features extracted on telephone bandwidth provide an assessment that is consistent with subjective intelligibility scores. Furthermore, the present TTS study is consistent with the earlier ASR study, in which a TTS system was used for template generation [6].

A next step is to evaluate the proposed approach against subjective intelligibility scores for speech degraded in telephony conditions, with further degradation types, such as background noises, noise suppression and various bit error patterns. With enough training data, a regression from average DTW distance to predicted Word Error Rate could be derived. Future work will also focus on the approach itself:

- We investigated speech intelligibility assessment at the sentence level. The approach could be extended to word level assessment, where Word Error Rate (WER) is estimated without performing ASR, using the utterance verification approach proposed in [21].
- Listeners have more than one "internal reference" for recognition. The approach could thus benefit from using more than one reference speech recording (similar to the template-based ASR system in [6, 12]), or from replacing reference speech by a statistical model such as Kullback-Leibler divergence-based HMM, which models the lexical content [24].
- The MLP in this study was trained to classify English phonemes. This makes the approach somewhat language and resource dependent. These issues could be addressed using an ANN that classifies multilingual phones, as done in the case of ASR, see e.g. [25, 26].

7. REFERENCES

[1] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, 1980.

[2] ——, "Mutual dependence of the octave-band weights in predicting speech intelligibility," *Speech Commun.*, vol. 28, no. 2, pp. 109–123, 1999.

[3] ITU-T Rec. P.862, *Perceptual evaluation of speech quality (PESQ)*. International Telecommunication Union, Geneva, Switzerland, 2001.

[4] 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.

[5] M. Cernak, A. Lazaridis, P. N. Garner, and P. Motlicek, "Stress and Accent Transmission In HMM-Based Syllable-Context Very Low Bit Rate Speech Coding," in *Proc. Interspeech*, 2014, pp. 2799–2803.

[6] S. Soldo, M. Magimai.-Doss, and H. Bourlard, "Synthetic References for Template-based ASR using Posterior Features," in *Proc. Interspeech*, 2012.

[7] 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.

[8] A. Hines and N. Harte, "Speech intelligibility prediction using a Neurogram Similarity Index Measure," *Speech Commun.*, vol. 54, no. 2, pp. 306–320, 2012.

[9] Y. Teng, R. Kubichek, R. Anderson-Sprecher, and J. E. Schroeder, "Objective Speech Intelligibility Measure for Low Bit-Rate Speech Codecs Operating in Noisy Channels," in *IEEE Mil. Commun. Conf.*, 2007, pp. 1–7.

[10] C. Middag, G. Van Nuffelen, J.-P. Martens, and M. De Bodt, "Objective intelligibility assessment of pathological speakers," in *Proc. Interspeech*, 2008, pp. 1745–1748.

[11] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on ASSP*, vol. 26, no. 1, pp. 43–49, 1978.

[12] S. Soldo, M. Magimai.-Doss, J. P. Pinto, and H. Bourlard, "Posterior features for template-based ASR," in *Proc. ICASSP*, 2011, pp. 4864–4867.

[13] ETSI TS 126 090, *Adaptive Multi-Rate (AMR) speech codec; Transcoding functions*. European Telecommunications Standards Institute, 2012.

[14] 3GPP2 C.S0014-E, *Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70, 73 and 77 for Wideband Spread Spectrum Digital Systems*. 3GPP2, 2011.

[15] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: the new Federal Standard at 2400 bps," in *Proc. ICASSP*, 1997, pp. 1591–1594.

[16] ITU-T Rec. P.501, *Test signals for use in telephonometry*. International Telecommunication Union, Geneva, Switzerland, 2012.

[17] ITU-T Rec. P.48, *Specification for an Intermediate Reference System*. International Telecommunication Union, Geneva, Switzerland, 1988.

[18] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Proc. Blizzard Challenge Workshop*, 2011.

[19] 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.

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

[21] M. Magimai.-Doss, G. Quer, R. Rasipuram, and M. Razavi, "An Information-Theoretic Edit Distance Formulation for Matching an Acoustic Speech Signal with a Text Hypothesis and its Potential Implications," Idiap, Tech. Rep. Idiap-Internal-RR-12-2015, 2015.

[22] ITU-T Rec. P.863, *Perceptual objective listening quality assessment (POLQA)*. International Telecommunication Union, Geneva, Switzerland, 2014.

[23] S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech Quality Estimation: Models and Trends," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 18–28, 2011.

[24] G. Aradilla, H. Bourlard, and M. Magimai.-Doss, "Using KL-based Acoustic Models in a Large Vocabulary Recognition Task," in *Proc. Interspeech*, 2008.

[25] D. Imseng, J. Dines, P. Motlicek, P. N. Garner, and H. Bourlard, "Comparing different acoustic modeling techniques for multilingual boosting," in *Proc. Interspeech*, 2012.

[26] R. Rasipuram and M. Magimai.-Doss, "Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model," Idiap, Tech. Rep. Idiap-RR-02-2014, 2014.