IDIAP RESEARCH REPORT

PROBABILISTIC SYMBOL SEQUENCE MATCHING AND ITS APPLICATION TO PATHOLOGICAL SPEECH INTELLIGIBILITY ASSESSMENT

Julian Fritsch Guillem Quer
Mathew Magimai.-Doss

Idiap-RR-01-2021

JANUARY 2021

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Probabilistic Symbol Sequence Matching and its Application to Pathological Speech Intelligibility Assessment

Julian Fritsch[1,2]*, Guillem Quer Romeo*[3]*, Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École polytechnique fédérale de Lausanne (EPFL), Switzerland
[3]SDG Group, Barcelona, Spain
{julian.fritsch, mathew}@idiap.ch

Abstract

Matching of a test signal to a reference word hypothesis forms the core of many speech processing problems, including objective speech intelligibility assessment. This paper first shows that the comparison of two speech signals can be formulated as matching of two sequences of "uncertain" or probabilistic latent symbols, in the same manner as string matching. Based upon that, we propose a pathological speech intelligibility assessment approach that compares pathological speaker's speech to control speaker's speech in phone space and articulatory feature space, and yields a score that is interpretable w.r.t. human listening test. Experimental validation of the proposed approach on the UA-speech corpus yielded a Spearman's correlation coefficient of 0.976 and a Pearson's correlation coefficient of 0.946.

Index Terms: String matching, Posterior features, Objective intelligibility Assessment, Pathological speech.

1. Introduction

Matching a reference word hypothesis with a test speech signal forms the core of many speech processing problems that focus on the message component in the speech signal, such as speech recognition, keyword spotting, speech intelligibility assessment. The reference word hypothesis can represented as: (a) speech signal(s) or (b) text. Depending on the choice of representation, we can broadly group the approaches to match a reference word hypothesis with a test speech signal into: (i) instance- or template-based approaches e.g., dynamic time warping based approaches [1] and (ii) statistical sequence model-based approaches e.g., hidden Markov model based approaches [2].

In recent years, posterior feature-based approaches for speech recognition [3–6] and speech assessment [7–9] have emerged, where class conditional probabilities of phones or articulatory features [10] are used as features. An interesting aspect is that in the posterior feature space, the instance-based and statistical sequence based approaches converge to a single formalism, where in both cases sequence of posterior probabilities are matched. Although posterior feature-based approaches have led to interesting developments, there is a gap in understanding what these approaches are doing in a theoretical sense. This paper is a step towards explaining that.

Towards that, in Section 2, we first elucidate that string matching can be interpreted as comparison of two sequences of categorical distributions. Based on that, we show that matching of two speech signals can be formulated as matching of two "uncertain" latent symbol sequences. Through theoretical links, we further show that the final matching score resulting from dynamic programming is an estimate of log-likelihood ratio, and thus has the same discrimination capabilities as edit distance in string matching. That is, a decision whether the reference word hypothesis and the test signal are same (word) or not can be made with low probability of error.

In Section 3, we validate the theoretical developments from Section 2 by applying it to pathological speech intelligibility assessment. More precisely, we develop an approach where impaired speech utterances of a speaker are compared to control i.e. healthy speakers' speech in phone space and articulatory feature space, and speaker intelligibility is measured as the percentage of correctness, similar to human listening tests. We demonstrate the effectiveness of the approach through a study on UA-speech corpus.

2. Probabilistic symbol sequence matching

This section establishes a link between string matching and matching of two speech signals.

2.1. Probabilistic interpretation of string matching

In computer science, it is well known that two strings that share the same symbol set can be matched by computing string edit distance [11]. One of the most commonly used string edit distances is Levenshtein distance [12], where the permissible edits are deletion of symbols, insertion of symbols and substitution of symbols. Given two strings $E = (e_1, \cdots e_m, \cdots e_M)$ and $O = (o_1, \cdots o_n, \cdots o_N)$ of lengths M and N, respectively, Levenshtein distance can be computed using dynamic programming as follows [11, 13]:

$$L(m,n) = \begin{cases} \max(m,n) & \text{if } \min(m,n) = 0 \\ \min \begin{cases} L(m-1,n)+1 \\ L(m,n-1)+1 \\ L(m-1,n-1)+1_{e_m \neq o_n}, \end{cases} & \text{else} \end{cases} \tag{1}$$

where e_m and o_n are categorical variables, $\{1, \cdots D\}$ is the set of categorical values or symbols in this case, $m \in \{0, 1, \cdots M\}$, $n \in \{0, 1, \cdots N\}$ and $1_{e_m \neq o_n} = 1$ if $e_m \neq o_n$ else 0. $L(M, N)$ yields the Levenshtein distance between the strings E and O.

The string edit distance computation can also be formulated in an information-theoretic manner as a comparison of two sequences of D dimensional categorical distributions. In other words, by treating e_m and o_n as categorical random variables. To illustrate that, the matching of string ABCD and string ACCDE is considered: If we use the English alphabet, then $D = 26$. The Levenshtein distance between the two strings, as shown in Figure 1, can be equivalently computed by comparing two sequences of 26-dimensional categorical distributions.

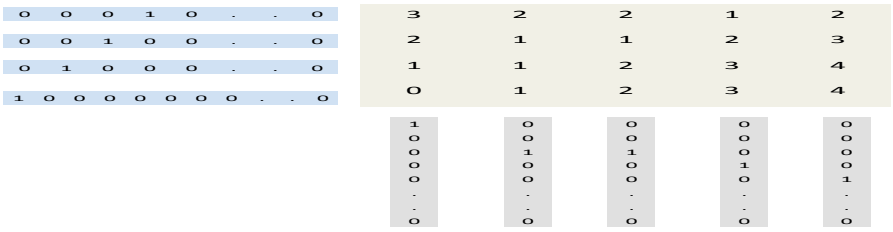

Figure 1: *Illustration of Levenshtein distance as comparison of sequence of categorical distributions.*

In this case, e_m and o_n are described by categorical distributions $\mathbf{y}_m = [P(e_m = 1) \cdots P(e_m = d) \cdots P(e_m = D)]^{\mathrm{T}}$ and $\mathbf{z}_n = [P(o_n = 1) \cdots P(o_n = d) \cdots P(o_n = D)]^{\mathrm{T}}$, respectively. Since there is no "uncertainty" about the symbols in both the sequences, the categorical distributions are Kronecker delta distribution or 1-of-D encoding (probability distributions with zero entropy). This yields two sequences of categorical distributions $Y = (\mathbf{y}_1, \cdots \mathbf{y}_m, \cdots \mathbf{y}_M)$ and $Z = (\mathbf{z}_1, \cdots \mathbf{z}_n, \cdots \mathbf{z}_N)$. These two sequences can be matched using Eqn. (1), where the condition $1_{\mathbf{e}_m \neq \mathbf{o}_n}$ is evaluated by hypothesis testing (i.e. two random categorical variables $\mathbf{e}_m$ and $\mathbf{o}_n$ belong to the same category/symbol or not) using a measure of discrimination such as Bhattacharya distance [14–16], Kullback-Leibler (KL) divergence [17, 18] to compare $\mathbf{y}_m$ and $\mathbf{z}_n$ and then thresholding the result. Formally,

$$1_{\mathbf{e}_m \neq \mathbf{o}_n} = \begin{cases} 0 & \text{if } l(\mathbf{y}_m, \mathbf{z}_n) \gtreqless \Delta \\ 1 & \text{else,} \end{cases} \tag{2}$$

where $l(\mathbf{y}_m, \mathbf{z}_n)$ denotes a measure of discrimination, Δ denotes a threshold and $\gtreqless$ is the condition chosen to make a decision. For example, if local score $l(,)$ is KL-divergence then $\Delta = 0$ and the condition to decide is $=$, as KL-divergence between Kronecker delta distributions is either 0 (matching perfectly) or ∞ (not matching). It can be shown that the same threshold and criteria are applicable to Bhattacharya distance, too.

2.2. Matching of speech utterances through comparison of uncertain latent symbol sequences

We can extend the probabilistic interpretation of string matching to matching of two speech utterances through latent symbols (e.g. phones) and verify whether those two utterances correspond to the same linguistic unit (e.g., word) or not. Let us suppose E^{sp} and O^{sp} represent two speech signals. We can estimate a match between them by,

1. Defining a latent symbol set $\{a^1, \cdots a^D\}$;
2. Splitting the speech signal E_{sp} into frames to yield a sequence of speech frames $(e_1^{sp}, \cdots e_m^{sp}, \cdots e_M^{sp})$, and then for each frame $m \in \{1, \cdots M\}$ estimating $\mathbf{y}_m^{sp} = [P(a^1|e_m^{sp}) \cdots P(a^D|e_m^{sp})]^{\mathrm{T}}$. Finally, yielding sequence of categorical distributions $Y^{sp} = (\mathbf{y}_1^{sp}, \cdots \mathbf{y}_M^{sp})$;
3. Splitting the speech signal O_{sp} into frames to yield a sequence of speech frames $(o_1^{sp}, \cdots o_n^{sp}, \cdots o_N^{sp})$, and then for each frame $n \in \{1, \cdots N\}$ estimating $\mathbf{z}_n^{sp} = [P(a^1|o_n^{sp}) \cdots P(a^D|o_n^{sp})]^{\mathrm{T}}$. Finally yielding sequence of categorical distributions $Z^{sp} = (\mathbf{z}_1^{sp}, \cdots \mathbf{z}_N^{sp})$; and
4. Matching the two sequences of categorical distributions Y^{sp} and Z^{sp}.

Matching Y^{sp} and Z^{sp} by applying Eqns. (1) and (2) would lead to a sub-optimal solution. As, unlike string matching, here we have "uncertain" or probabilistic symbols. In other words, the latent symbols are not directly observable from the speech signal. We can only get a probabilistic estimate of the latent symbols leading to $\mathbf{y}_m^{sp}$ and $\mathbf{z}_n^{sp}$ with non-zero entropies, $\forall m$ and $\forall n$. Computation of $1_{\mathbf{e}_m^{\mathbf{sp}} \neq \mathbf{o}_n^{\mathbf{sp}}}$ amounts to early decision making, which can be error prone. The alternative is to delay the decision, i.e. estimate $l(\mathbf{y}_m^{sp}, \mathbf{z}_n^{sp})$ through a measure of discrimination, e.g. KL-divergence, Bhattacharya distance and to not make a decision. Under this condition, we can rewrite the dynamic programming in Eqn. (1) as,

$$L^{sp}(m,n) = \begin{cases} 0 & \text{if } m = 0 \text{ and } n = 0 \\ \infty & \text{else if } \min(m,n) = 0 \\ l(\mathbf{y}_m^{sp}, \mathbf{z}_n^{sp}) + \min[L^{sp}(m-1,n), \\ L^{sp}(m,n-1), L^{sp}(m-1,n-1)] & \text{else.} \end{cases} \tag{3}$$

The match between E^{sp} and O^{sp} is finally given by $L^{sp}(M,N)$.

Although "local" decision making is skipped, $L^{sp}(M,N)$ would still have the same properties of $L(M,N)$ in string matching. The reason being that KL-divergence [18] [19, Chapter 4], in more general sense J-divergence, Bhattacharya distance [15] are functions of likelihood ratio. To be more precise, they can be interpreted or shown as an estimate of log-likelihood ratio. If we consider that, then $L^{sp}(M,N)$ is nothing but a sum of log-likelihood ratios. So, $L^{sp}(M,N)$ can be regarded as a probabilistic or soft edit cost to transform latent symbol sequence corresponding to E^{sp} into latent symbol sequence corresponding to O^{sp} and vice versa, like edit

distance in string matching is a cost to transform one string into another string. This leads to the hypothesis that, through $L^{sp}(M, N)$, we should be able to decide with a low probability of error whether E^{sp} and O^{sp} correspond to the same linguistic unit (e.g., word) or not.

2.3. Validation study

We conducted an utterance verification study on Phonebook corpus [20] to validate the hypothesis presented in the previous section. We adapted the 600 words speaker-independent task-independent ASR task [21]. The terms speaker-independent task-independent mean that the speakers and the words in the training set, validation set and test set are entirely different. Table 1 provides an overview.

Table 1: *Overview of the PhoneBook corpus.*

Number of	Train	Cross-validation	Test
Utterances	19421	7290	6598
Speakers	243	106	96
Words	1580	603	600

We used a multilayer perceptron to estimate $\mathbf{y}_m^{sp}$ and $\mathbf{z}_n^{sp}$: A five layer MLP was trained to classify the latent symbols, i.e. 42 context-independent (CI) phonemes. The input to the MLP was 39 dimensional PLP cepstral coefficients with four frame preceding and four frame following context (39×9). The features were extracted using HTK with a frame size of 25 ms and a frame shift of 10 ms. The MLP was trained with cross entropy cost function using the Quicknet tool [22].

For the utterance verification task, we created (a) 150K positive pairs (i.e. pair of utterances containing same word) and (b) 150K negative pairs (i.e. pair of utterances containing different words) on the test set data. We estimated *path length normalized* $L^{sp}(M, N)$ for each pair of utterances using symmetric KL-divergence (SKL) as the local score $l(\mathbf{y}_m^{sp}, \mathbf{z}_n^{sp})$, and computed the area under curve (AUC) of the receiver operator curve. The AUC is 0.998. This indicates that indeed with $L^{sp}(M, N)$ we can decide whether two speech utterances are the same word or not with low probability of error.

3. Application to Pathological Speech Intelligibility Assessment

A way to assess speech intelligibility of speakers with speech impairment such as dysarthria is to ask the speakers to produce a set of words; perform a human listening test, and assess intelligibility in terms of percentage of recognized words. We could emulate such an intelligibility assessment in an objective manner by building upon the discrimination capabilities of $L^{sp}(M, N)$ in the following way:

1. Replace the human listeners by collecting utterances of those words from a set of control i.e. healthy speakers.
2. For each word, compare the speech utterance of the impaired speaker produced with each one of the control speakers utterance, as described in Section 2.2, and take a majority vote based on $L^{sp}(M, N)$ to decide if the speaker produced the word correctly or not.
3. Compute the percentage of correctly produced words.

We investigate this approach through a study on the Universal Access (UA) speech database.

3.1. Database

We conduct our experiments on the UA speech database [23], consisting of 15 English-speaking cerebral palsy patients (11 males, 4 females) and 13 healthy speakers (9 males, 4 females). Each speaker has uttered 765 isolated words, 155 isolated words, repeated 3 times, the remaining 300 spoken only once. The subjective intelligibility scores of patients range from 2% to 95%. We consider the recordings of the 5th channel for our evaluation. Furthermore, an energy-based voice activity detection using Praat ([24]) is used to extract speech segments only. Each subjects intelligibility score was obtained by letting five naive listeners transcribe the isolated words and then calculating the average number of correct transcriptions.

3.2. Latent symbol space and $\mathbf{y}_m^{sp}$ and $\mathbf{z}_n^{sp}$ estimators

We considered two different latent symbol spaces, namely, phone space and articulatory feature space.

Phone space: We used 45 context-independent phonemes in UniSyn dictionary. We used an off-the-shelf multilayer perceptron (MLP). The MLP is trained on 232 hours of conversational telephone speech to classify 44 English phonemes and silence, i.e., $K = 45$ output units. The MLP inputs are 39-dimensional perceptual linear predictive cepstral features with (frame size is 25 ms, frame shift 10 ms) a nine frame temporal context (i.e., four frames preceding and four frames following). The MLP was trained with the QuickNet tool [22] by minimizing the frame-level cross entropy.

Articulatory feature (AF) space: There are different ways to represent phonemes as articulatory features such as binary features [25] or multi-valued features [26]. In this work, we conducted studies with binary features and multi-valued AF representations:

- AF_{binary}: The Phonet toolkit [27] was used to extract binary AFs attributed to the manner of articulation. Phonet consists of 18 recurrent neural network-based binary classifiers. These classifiers have been trained on 17 hours of clean FM podcasts in Mexican Spanish.
- AF_{multi}: We used an off-the-shelf CNN-based multi-valued AF features trained on AMI corpus with raw waveform as input [28]. The multi-valued articulatory features were based on the previous work on automatic speech recognition [10]. There are four CNNs corresponding to articulatory feature $f \in \{\texttt{Manner}, \texttt{Place}, \texttt{Height}, \texttt{Vowel}\}$.

In this case, $\mathbf{y}_m^{sp} = [\mathbf{y}_{m,1}^{sp} \cdots \mathbf{y}_{m,f}^{sp} \cdots \mathbf{y}_{m,F}^{sp}]$ and $\mathbf{z}_m^{sp} = [\mathbf{z}_{n,1}^{sp} \cdots \mathbf{z}_{n,f}^{sp} \cdots \mathbf{z}_{n,F}^{sp}]$ are stack of AF posterior distributions. $F = 18$ for AF_{binary} and $F = 4$ for AF_{multi}.

3.3. Objective score estimation

To estimate dysarthric speaker's speech intelligibility, we compute an objective score in the following manner:

1. For each word in the list, match the dysarthric speaker's utterance of that word with each of the control speaker's utterance of that word as per (3). If the number of control speakers is C then this will yield C $L_c^{sp}(M_c, N_{dys})$ scores. M_c denotes the number of frames in control speaker c's speech utterance. N_{dys} denotes the number of frames in dysarthric speaker's speech utterance.
2. Get a majority vote based on the decision made with all of the $L_c^{sp}(M_c, N_{dys})$ scores that the utterance of dysarthric speaker and control speaker are the same word or not (utterance verification). The underlying hypothesis is that the higher the score $L_c^{sp}(M_c, N_{dys})$ is the more the dysarthric speech deviates from the healthy speakers' pronunciation of the word, thus the word is less intelligible.
3. Given the majority vote for each word, get the percentage of words correctly spoken as the objective intelligibility score for the dysarthric speaker.

In order to decide whether the utterance of a dysarthric speaker and control speaker are the same word or not based on path length normalized $L_c^{sp}(M_c, N_{dys})$. We investigate two different methods:

1. **UV-based**: From the control speakers utterances, we created same word utterance pairs and different word utterance pairs; matching each of those pairs according to (3) and obtaining a histogram of match score for "same word" and a histogram of match score for "different word". To demonstrate the effectiveness of the approach, we created two decision boundaries: (i) one at the intersection of the histograms as illustrated in Fig. 2, referred to as Thr_{inter} and (ii) one at the center of the two means of the histogram, referred to as Thr_{cen}. The decision criteria was $L_c(M_c, N_{dys}) \leq Thr_{inter}$ and $L_c(M_c, N_{dys}) \leq Thr_{cen}$ to make "same word" decision, respectively.

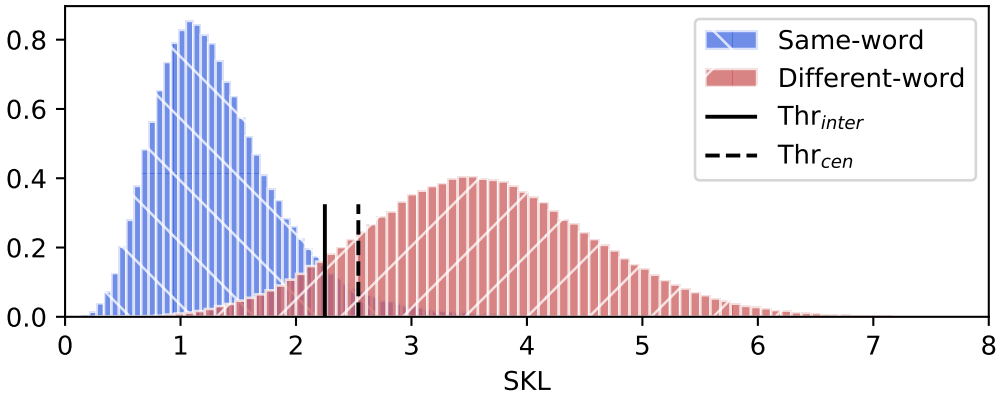

Figure 2: *Distribution of same-and different-word scores using SKL as cost function.*

2. **Posterior estimate-based**: As discussed earlier, $L^{sp}(M, N)$ can be regarded as an estimate of log-likelihood ratio. In such a case, we should be able to decide by estimating the posterior probability of being the same word either as,

$$P_c(\texttt{sw}|E_c^{sp}, O_{dys}^{sp}) = \frac{2}{1 + \exp\left(L_c^{sp}(M_c, N_{dys})\right)}, \tag{4}$$

or as

$$P_c^{\alpha}(\texttt{sw}|E_c^{sp}, O_{dys}^{sp}) = \frac{2}{1 + \exp\left(L_c^{sp}(M_c, N_{dys}) - \alpha\right)}, \tag{5}$$

where sw denotes the class "same word", E_c^{sp} control speaker's speech, O_{dys}^{sp} denotes dysarthric speaker's speech and α is an offset value.

Eqn. (4) presumes that, for probability of 1, $L_c^{sp}(M_c, N_{dys})$ should go to zero (i.e. perfect match). This happens in string matching as there is no ambiguity or uncertainty about the symbols. However, in the case of matching speech utterances, we are dealing with "uncertain" or probabilistic symbols, as a consequence obtaining a perfect match is highly improbable. α in Eqn. (5) tends to offset that bias or effect. α can be based on the "same word" match score obtained on control speakers data in UV-based approach. In this work, we simply took the minimum of those scores as α value.

The decision criteria then simply can be $P_c(\texttt{sw}|E_c^{sp}, O_{dys}^{sp}) \geq 0.5$ and $P_c^{\alpha}(\texttt{sw}|E_c^{sp}, O_{dys}^{sp}) \geq 0.5$, respectively.

Note that in all cases we are dealing with path length normalized $L_c^{sp}(M_c, N_{dys})$ scores. Furthermore, we use SKL as the local score $l(\mathbf{y}_m^{sp}, \mathbf{z}_n^{sp})$.

3.4. Results and analysis

Table 2 shows the results obtained for the two proposed approaches: UV-based and posterior estimate-based in terms of Spearman's correlation coefficient ρ and Pearson's correlation coefficient r. AF$_{multi-manner}$ is the case where only manner of articulation of CNN

posteriors are used. Within the posterior estimate-based approach, we do not report for AF_{binary} and AF_{multi}, as it requires a new implementation to estimate posterior probability for each AF f independently and combining them. This is part of our future work. We also present results reported in the literature on the same data set. It can be observed that both UV-based and posterior estimate-based approach are yielding high correlations, and perform comparable to or better than the approaches reported in the literature. In terms of latent symbol space, phone space yields better intelligibility assessment than AF space. When comparing across AF space, AF_{binary} which is based on manner of articulation performs better than AF_{multi}. If we only consider manner of articulation, i.e. $\text{AF}_{multi-manner}$, the performances are comparable. We did not report p-values due to space constraint. The maximum p-value obtained by the proposed systems is 0.0009. This indicates that our results are statistically significant.

Table 2: *Pearson's correlation and Spearman's correlation between subjective and objective intelligibility.*

Latent symbol space	UV-based			
	Thr_{cen}		Thr_{inter}	
	ρ	r	ρ	r
Phone	.909	.920	.957	**.946**
AF_{binary}	.885	.914	.885	.919
AF_{multi}	.806	.762	.819	.763
$\text{AF}_{multi-manner}$	.894	.911	.885	.893
	Posterior estimate-based			
	P_c		P_c^α	
Phone	.956	.831	**.974**	.902
$\text{AF}_{multi-manner}$	.860	.787	.894	.914
Baseline systems				
P-ESTOI [29]	.94	.94		
iVectors [30]	-	.91		
Discriminant analysis [31]	-	.92		
Temporal dynamics [32]	0.85	0.87		

Figure 3 shows the Pearson's correlation plot overlaid for different systems. It can be observed that the UV-based approach with Phone space is predicting percentage correctness close to subjective correct percentage score in high intelligibility regions, while AF_{binary} is predicting well very low intelligibility regions. This suggests that combining Phone space and AF space could have added benefits. The posterior estimate-based approach is under-predicting the subjective correct percentage score. This issue can be addressed by calibrating α or the threshold.

4. Conclusions

In this paper, we elucidated that string matching can be interpreted as a comparison of two sequences of categorical distributions. Through that interpretation, we showed that matching two speech signals by comparing their sequences of class conditional probabilities of latent symbols (e.g. phones, articulatory features) is equivalent to probabilistic string matching. Thus, the resulting match score has similar discrimination capabilities to edit distance in string matching. We experimentally validated that through (a) an utterance verification study and (b) through the development of a pathological speech intelligibility assessment approach that emulates a human subjective listening test for intelligibility assessment.

Although, we have used neural networks to estimate posterior probabilities of latent symbols ($\mathbf{y}_m^{sp}$ and $\mathbf{z}_n^{sp}$), the developments in Section 2 are applicable to other posterior probability estimators as well. In the same vein, there are several local scores other than SKL that can be used to match sequences of posterior probabilities [33]. Finally, the developments in this paper apply equally to KL-HMM [3, 4]. Our future work will build upon that for pathological speech assessment.

5. Acknowledgements

This work was funded through European Union's Horizon H2020 Marie Skłodowska-Curie Actions Innovative Training Network European Training Network (MSCA-ITN-ETN) project TAPAS under grant agreement No. 766287. The second author's contribution lies in the work presented in Section 2, which was conducted within the scope of his Master's thesis research [34], as an intern at Idiap between May-October 2014 under the supervision of the third author.

6. References

[1] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 1, pp. 43–49, February 1978.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

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

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

[5] G. Aradilla and H. Bourlard, "Posterior-based features and distances in template matching for speech recognition," in *4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, 2007.

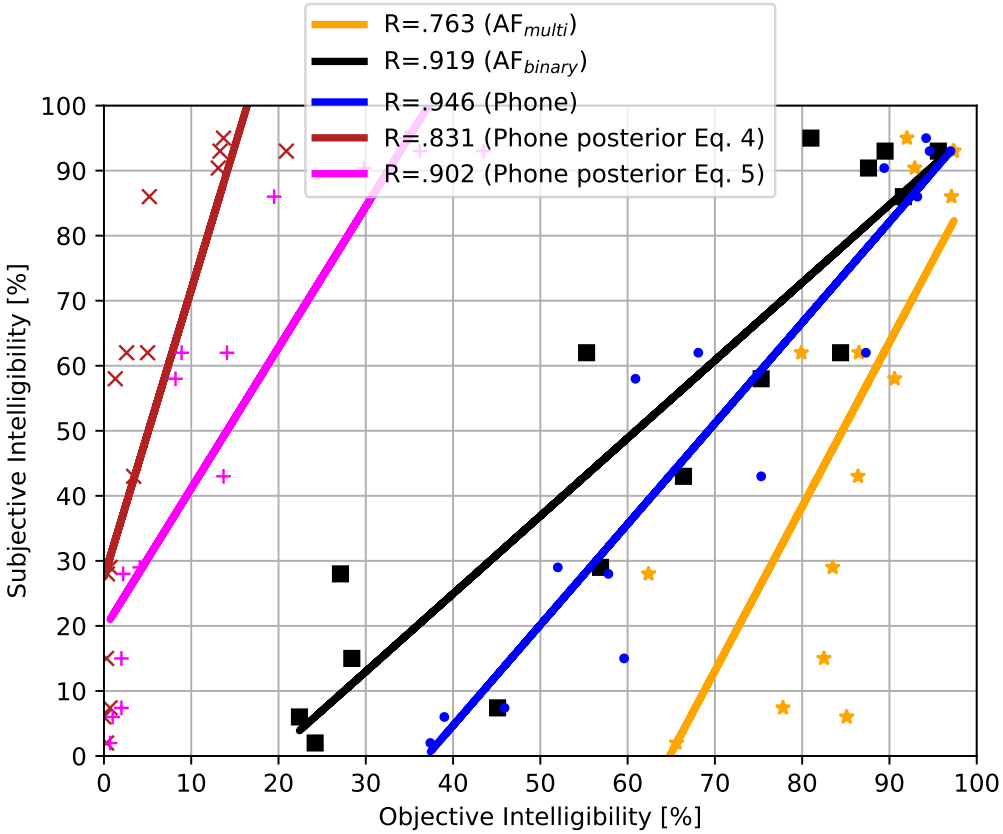

Figure 3: *Pearson's correlation plot obtained from proposed intelligibility assessment systems.*

[6] S. Soldo, M. Magimai.-Doss, J. Pinto, and H. Bourlard, "Posterior Features for Template-based ASR," in *Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2011, pp. 4864–4867.

[7] R. Ullmann, M. Magimai.-Doss, and H. Bourlard, "Objective speech intelligibility assessment through comparison of phoneme class conditional probability sequences," in *Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2015.

[8] R. Ullmann, R. Rasipuram, M. Magimai.-Doss, and H. Bourlard, "Objective intelligibility assessment of text-to-speech systems through utterance verification," in *Proceedings of Interspeech*, 2015.

[9] R. Rasipuram, M. Cernak, A. Nanchen, and M. Magimai.-Doss, "Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities," in *Proceedings of Interspeech*, 2015.

[10] R. Rasipuram and M. Magimai.-Doss, "Articulatory feature based continuous speech recognition using probabilistic lexical modeling," *Computer Speech & Language*, vol. 36, pp. 233–259, 2016.

[11] D. Sankoff and J. Kruskal, "Time warps, string edits and macromolecules: The theory and practise of sequence comparison," in *CSLI Publications, Leland Stanford Junior University*, 1999.

[12] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[13] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of ACM*, vol. 21, no. 1, pp. 168–173, 1974.

[14] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[15] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun.*, vol. COM-15, no. 1, pp. 52–60, 1967.

[16] N. A. Thacker, F. J. Aherne, and P. I. Rockett, "The Bhattacharyya metric as an absolute similarity measure for frequency coded data," *Kybernetika*, vol. 34, no. 4, pp. 363–368, 1997.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[18] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. on Information Theory*, vol. IT-20, no. 4, pp. 405–417, 1974.

[19] R. Blahut, "Principles and practice of information theory," in *Addison-Wesley*, 1987.

[20] J. F. Pitrelli, C. Fong, S. H. Wong, J. R. Spitz, and H. C. Leung, "PhoneBook: A phonetically-rich isolated-word telephone-speech database," in *Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1995, pp. 1767–1770.

[21] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite, "Hybrid HMM/ANN systems for training independent tasks: Experiments on 'PhoneBook' and related improvements," in *Proceedings of IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1997, pp. 1767–1770.

[22] D. Johnson, D. Ellis, C. Oei, C. Wooters, and P. Faerber, "Quicknet," *www1.icsi.berkeley.edu/Speech/qn.html*, 2004.

[23] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[24] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[25] N. Chomsky and M. Halle, *The Sound Patterns in English*. MIT Press, 1968.

[26] P. Ladefoged, *A Course in Phonetics*. Harcourt Brace College Publishers, 1993.

[27] J. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, "Phonet: a tool based on gated recurrent neural networks to extract phonological posteriors from speech," *Proceedings of Interspeech*, pp. 549–553, 2019.

[28] J. Fritsch, S. P. Dubagunta, and M. Magimai.-Doss, "Estimating the degree of sleepiness by integrating articulatory feature knowledge in raw waveform based cnns," in *Proceedings of ICASSP*, 2020, pp. 6534–6538.

[29] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Spectral subspace analysis for automatic assessment of pathological speech intelligibility," in *Proceedings of Interspeech*, 2019.

[30] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 3, pp. 1–21, 2015.

[31] M. S. Paja and T. H. Falk, "Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[32] T. H. Falk, R. Hummel, and W. Chan, "Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility," in *Proceedings of ICASSP*, 2011, pp. 4480–4483.

[33] S. Soldo, M. Magimai.-Doss, J. Pinto, and H. Bourlard, "On MLP-based Posterior Features for Template-based ASR," Idiap, Tech. Rep. Idiap-RR-37-2009, 2009.

[34] G. Quer Romeo, "An information-theoretic string matching approach for spoken utterance verification and keyword spotting," Master's thesis, Universitat Politècnica de Catalunya, 2015, https://www.recercat.cat/handle/2072/344287.