Study of Time and Frequency Variability in Pathological Speech and Error Reduction Methods for Automatic Speech Recognition

Oscar Saz, Antonio Miguel, Eduardo Lleida, Alfonso Ortega, Luis Buera

Communication Technologies Group (GTC)
Aragon Institute of Technology (I3A), University of Zaragoza, Spain
{oskarsaz,amiguel,lleida,ortega,lbuera}@unizar.es

Abstract

In this work, we study the variations in the time and frequency domains inside a Spanish language corpus of speakers with non-pathological and pathological speech. We show how pathological speech has a greater variability in the duration of the words than non-pathological speech, while in the frequency domain we show that the vowels confusability increases by a 18%. The baseline experiments in Automatic Speech Recognition (ASR) with this corpus demonstrate that this variability causes a loss in the performance of ASR systems. To reduce the impact of time and frequency variability we use a recent Vocal Tract Length Normalization (VTLN) system: MATE (augMented stAte space acousTic modEl), as a way of improving the performance of ASR systems when dealing with speakers who suffer any kind of speech pathology. Experiments with MATE show a 17.04% and 11.19% WER reduction by using frequency and time MATE respectively.

1. Introduction

Speech pathologies such as dysarthria, dyslalia, dysglossia or aphasia [1] affect dramatically the communication abilities of those who suffer them. Specially when suffered from a very young age, these pathologies have a very negative impact in the social development of the patients with any of them, which is a real manforce loss for our society. Altough the range of causes and symptoms of these pathologies is very wide, these speech handicaps can mainly be originated by one of these three reasons:

- Different ways of brain damage, like cerebral palsy or stroke, can lead to dysarthria or aphasia [1], where the articulatory abilities of the speaker are limited by the brain's disability for controlling the organs used in the speech articulation.
- When any of the organs of the articulatory system (tongue, mouth, vocal chords,...) is affected in its morphology or movement, it may lose its ability to generate correct speech. This situation leads to the presence of a dysglossia. The range of dysglossias depends on which one is the ill organ. When lungs cannot provide enough air for the emission of speech due to a pulmonar affection, it turns into a severe chronic aphony.
- Finally, when there is no physical disability that affects speech, and usually related to a type of mental retardation like Down Syndrome, a dyslalia can appear as the other main type of speech pathology. In this situation, the patient mistakes or misses differents sounds and phonemes during the production of the speech.

The study of all of these pathologies from a phonological point of view has clearly shown how they affect the normal production of speech, resulting on a, sometimes critical, variation of the main parameters of speech [2].

However, the use of speech technologies could be very useful for the patients of all these pathologies. Computer aided systems for speech therapy and for augmentative communication [3] have been studied and developed in the last years, becoming a very promising research line in these days. As shown in these works, the use of Automatic Speech Recognition (ASR) systems could really help and improve the quality of life of these patients. However, the phonological variations in speech due to these pathologies are a real challenge in the use of ASR systems, as long as the performance of the conventional systems is very variable and highly dependent on the conditions of the speaker [3].

This paper is organized as follows. In Section 2, the test corpus of Spanish pathological speech is introduced. In section 3 an analysis of the pathological speech in the time and frequency domains is made, and differences between pathological and non-pathological speech are studied. In section 4 MATE warping method is presented as an effective way of fighting against the time and frequency distortions in pathological speech. In section 5 the ASR results for the baseline and the improvements obtained with MATE are presented to finally extract the conclusions of this work in section 6.

2. The Corpus

The corpus used for this work was recorded by the Department of Signals and Communications from the University of Las Palmas de Gran Canaria (Spain). The corpus was originally used for the research in identification and classification of pathological speech [4].

The corpus has 3,547 utterances of different words spoken by speakers with several types of pathology, ranging from non-pathology to severe pathologies. Every utterance contains an isolated word from the set of possible words. There is also a wide variety of ages and genders among the speakers. All the speakers share the accent of the Canary islands, very characteristic in the range of Spanish accents.

2.1. Phonological content of the corpus

The corpus contains utterances of the 57 Spanish words that are included in the "Induced Phonological Register" [5]. This set of words contains a phonetically rich selection of words including

This work has been supported by the national project TIN 2005-08660-C04-01

nearly all the phonemes in the Spanish language; as well as dyptongues and other singularities of the language. The length of the words is balanced and ranging from 1-syllable words to 4-syllable words.

2.2. Speaker composition of the corpus

For the corpus, 19 reference speakers without pathology were recorded. The classification of these speakers in age and sex is:

- 2 males and 1 female in the range of 6-12 years old.
- 1 female in the range of 12-15 years old.
- 3 males and 3 females in the range of 15-35 years old.
- 2 males and 3 females in the range of 35-60 years old.
- 2 males and 2 females over 60 years old.

However, not all the utterances were kept for the final corpus, and 1,077 utterances of the 1,083 original utterances were used for this work.

The number of speakers for the study of pathological speech was 30, with the following classification in age and sex:

- 3 males and 2 females in the range of 6-12 years old.
- 2 males and 3 females in the range of 12-15 years old.
- 3 males and 2 females in the range of 15-35 years old.
- 5 males and 3 females in the range of 35-60 years old.
- 3 males and 4 females over 60 years old.

Due to the limitations of some of the speakers, not all the utterances were included in the ultimate corpus, which made the 1,710 original utterances get reduced to 1,615 utterances. This number of utterances were enlarged with 5 patients with severe chronic aphony, a male in the range of 15-35 years old and 4 males in the range of 35-60 years old. Each one of them uttered 3 times every one of the 57 words of the "Induced Phonological Register", which made the total number of utterances of pathological speech rise to 2,470 utterances.

3. Analysis of patological speech

In this section, we present how the speech signal differs from a speaker whitout any speech pathology to speakers with pathological speech. The analysis is made in time and frequency domains. From this study we try to understand which characteristics of the pathological speech affect the performance of the ASR systems when dealing with this kind of users.

For this purpose, we took advantage of the previous work made over this corpus [4], where an automatic phonetic segmentation and labelling was made on the two parts of the corpus, and the errors of the automatic segmentation were manually corrected to get a totally accurate segmentation into phonemes.

3.1. Time-domain analysis

For the time-domain analysis we computed the mean duration and the standard deviation, refered to as a percentage of the mean duration, of the 57 words across the corpus, obtaining the mean results for all the words with the same number of syllables, as it is shown in Tables 1 and 2.

Word syllables	Mean	Standard deviation
1	472.6 msec.	15.11%
2	542.4 msec.	16.56%
3	633.4 msec.	14.20%
4	762.4 msec.	14.98%

Table 1: *Mean duration and standard deviation for non-pathological speech.*

Word syllables	Mean duration	Standard deviation
1	427.8 msec.	26.77%
2	530.9 msec.	24.33%
3	641.8 msec.	26.62%
4	845.5 msec.	38.26%

Table 2: *Mean duration and standard deviation for pathological speech.*

This analysis shows that the difference between pathological and non-payhological speech is not the mean duration of the words, but the duration variability, measured by the standard deviation. The words uttered by non-pathological speakers have a standard deviation of 15% around the mean, while the words uttered by pathological speakers present a standard deviation of 25% around the mean, even 38% for 4-syllable words. These results are due to the large variety of pathologies in the corpus; some of the speakers tend to length the pronunciation of the different sounds, while some other speakers shorten sounds, even in some cases the phoneme is not pronounced by the speaker. For an ASR system based on Hidden Markov Models (HMM), the length of the sounds is a critical parameter in the performance of the system [9].

3.2. Frequency-domain anaylisis

For the frequency-domain analysis we studied the variations of the first and second formants ($F1$ and $F2$) of the five Spanish vowels ($/a/$, $/e/$, $/i/$, $/o/$, $/u/$) across the whole set of signals for non-pathological and pathological speech. For this purpose we processed the whole corpus with a 12-order Linear Predictive Coefficients (LPC) analysis, to obtain the mean and the standard deviation of the vowels considered as a two-dimensional Gaussian where the dimensions are $F1$ and $F2$ respectively. The formants obtained with the LPC analysis where manually reviewed to discard the prediction errors.

The results of the LPC analysis of the non-pathological part of the corpus are shown in Figure 1, where we present the usual distribution of the Spanish vowels in the $F1$-$F2$ space, plotting the samples that lay inside the ellipse with a 95% confidence interval for the learned $F1$-$F2$ model. This distribution is known as the formantic triangle of the vowels.

For the study of the diferences between pathological and non-pathological speech in the formants distribution, we computed the Kullback-Leibler distance for every possible pair of vowels. The Kullback-Leibler distance measures the distance between two probabilistic distributions, in this case, between two 2-dimensional Gaussian distributions of the two vowels: $A \sim \mathcal{N}(\mu_A, \Sigma_A)$ and $B \sim \mathcal{N}(\mu_B, \Sigma_B)$ where μ_A and μ_B are the mean vectors and Σ_A and Σ_B the diagonal covariance matrices. The Kullback-Leibler expression in this case is:

$$KL(A,B) = \sum_i (log(\frac{\Sigma_{A_i}}{\Sigma_{B_i}}) + \frac{(\mu_{A_i} - \mu_{B_i})^2}{\Sigma_{B_i}} + \frac{\Sigma_{A_i}}{\Sigma_{B_i}} - 1), \quad (1)$$

where i is every one of the dimensions of the Gaussian distribution.

However, from the definition of the Kullback-Leibler distance $KL(A,B) \neq KL(B,A)$, so we had to compute a modified Kullback-Leibler distance in order to provide the simmetry property to the distance expression:

$$KL_{modified}(A,B) = \frac{KL(A,B) + KL(B,A)}{2}. \quad (2)$$

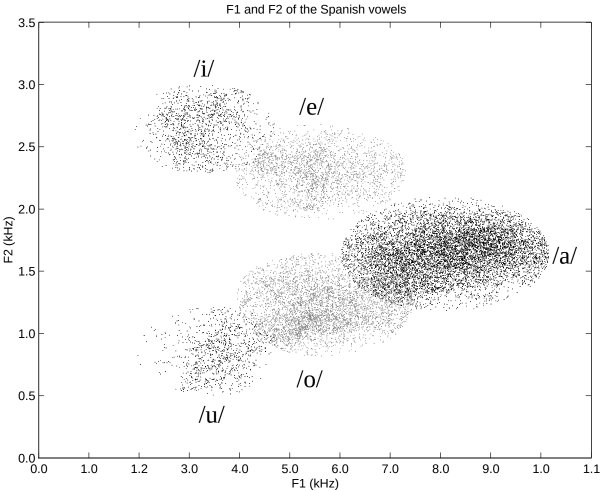

Figure 1: *$F1$ and $F2$ of Spanish vowels.*

Vowels	Non-path. speech	Path. speech	Decrease
$/a/ - /e/$	29.50	28.78	2.47%
$/a/ - /o/$	14.69	12.67	13.78%
$/e/ - /o/$	44.93	30.95	31.11%
$/e/ - /i/$	24.72	18.58	24.83%
$/o/ - /u/$	66.40	55.68	18.64%

Table 3: *Kullback-Leibler Distance.*

Table 3 shows the Kullback-Leibler distance between the most easily confused vowels in the $F1$-$F2$ vowel map. The results show clearly an important reduction in the average Kullback-Leibler distance (18.17%) for pathological speech, compared to the results for non-pathological speech. To verify this result, we also computed the Fisher's Ratio for the same couples of vowels. This Ratio gives a measure of the discriminating power of a couple of variables; and the results showed also an average decrease (19.14%) for pathological speech when compared to non-pathological speech. As a result of this, we can conclude that there is more confusability among the vowels in pathological speech, due to a variation in the frequency of the formants, and we can also point this as a major problem for ASR systems.

4. Warping methods

In order to compensate both sources of variability some methods have appeared previously as VTLN [6]. In this paper we focus the experiments towards pathological speech, which is found to have a broader range of vocal tract shape changes and articulatory variable speed. In [7], a method for on-line local reduction of the mismatch between data and model was presented. The model framework, MATE, implies an expansion of the VTLN methods which provides spectral warping either the dynamic feature computation to be locally optimized, simultaneously to the decoding of the state sequence.

MATE model has shown good performances in noise free or moderately noisy speech conditions [7] by tracking the vocal tract shape changes during speech utterances thanks to a new degree of freedom added to the standard HMM.

As it was shown in [8] the spectral warping performed in VTLN methods is equivalent to a linear projection of the cepstral feature space, we can find equivalent transformation matrices such us, $\mathbf{X}^{\alpha_n} = \mathbf{A}_n\mathbf{X}$, where $n = 1, \cdots, N$, is the index of the warping factor. The MATE model is constructed expanding the state space as in [7], where a state q is expanded into states (q, n) with n the transformations or warpings. The new model provides observation generation probability density functions (pdfs) in the states that depend on a discrete set of transformation matrices, $\{\mathbf{A}_n\}_{n=1}^{N}$, embedding the warping in the model as a general transformation.

Given that a component in the original state q pdf mixture follows normal distribution: $\mathcal{N}(\mu_q, \mathbf{\Sigma}_q)$, the expanded state components are then assumed to follow a distribution:

$$\mathbf{x}_t|_{n,q} \sim \mathcal{N}(\mathbf{A}_n\mu_q, \mathbf{A}_n\mathbf{\Sigma}_q\mathbf{A}_n^t), \tag{3}$$

so that the model can generate sequences warped cepstrum vectors, which are expected to be closer to real data. In the expanded state space, the transition probabilities follow a Multinomial distribution of parameters:

$$\mathbf{\Pi} = \{\pi_{q',n',q,n}\}_{q'=1,n'=1,q=1,n=1}^{Q,N,Q,N}, \tag{4}$$

being $\pi_{q',n',q,n}$ the transition from state (q', n') to (q, n) probability, with the constraint $\sum_{q,n} \pi_{q',n',q,n} = 1, \forall q', n'$. The complete parameter set for the MATE model is composed by $\mathbf{\Pi}$ and the state *pdfs* described in 3.

The search algorithm for decoding unlabeled sequences under this framework will be given by the recursive equation:

$$\phi_{q,n}(t) = \max_{n',q'} \{\phi_{q',n'}(t-1) \cdot \pi_{q',n',q,n}\} \cdot f(\mathbf{x}_t|q, n), \tag{5}$$

where $\phi()$ is the score state variable and $\mathbf{\Pi}$ vector contains the state transition probabilities and $f(\mathbf{x}_t|n, q)$ is the observation generation pdf described in (3).

This recursive expression is very similar to the one in [7] being the main difference how the warping is done, since now is the model who tries to generate or evaluate the warped data instead of normalizing data to fit the model. In this framework the covariance is normalized in the model description so the Jacobian normalization in [8] is included in the model. The time warping method used for the experiments of this work is the same as explained in [7].

5. Results

Once the corpus has been presented and some analyses of the variability in the time and frequency domain have been made, the experiments performed for testing the feasibility of time and frequency warping methods for the ASR of pathological speech will be presented in this section.

Due to the small amount of data for the task of speech recognition in this corpus, we created four subtasks with cross-data of train and test. The results we present in this work are the mean and standard deviation (STDV) in WER (Word Error Rate) of the results in this subtasks for both non-pathological and pathological speech.

For this work, the signals were resampled to 16 kHz from the 22.05 kHz original sampling frequency. We used a window length of 25 ms. with a window shift of 10 ms. Our parametrization was a 39 MFCC (Mel-Frequency Cepstrum Coefficients) parametrization, with 12 static parameters, 12 delta-parameters and 12 delta-delta-parameters plus the energy, delta-energy and delta-delta-energy parameters.

The use of HMM for ASR has been widely accepted during decades, but when dealing with pathological speech the election of the HMM structure does not necessarily have to agree with

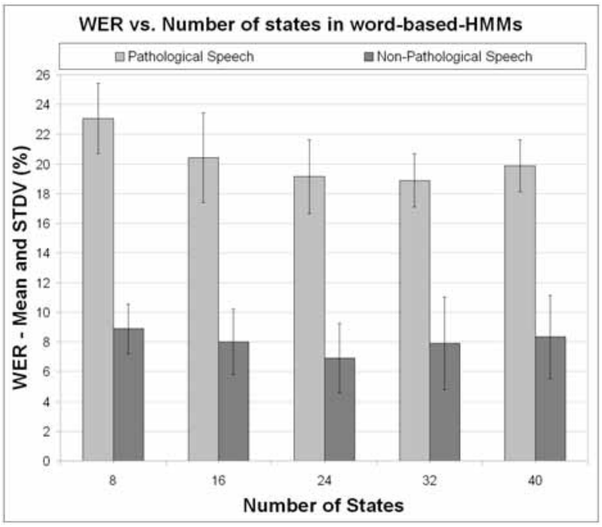

Figure 2: *Word Error Rate for different number of states.*

the traditional structures [9]. Therefore we decided to make a previous study for correctly selecting the HMM structure. Word models were used in all the experiments. In order to study how the number of states in our word-based HMM affected the performance of the ASR system, an experiment was performed and the results can be seen in Figure 2, where it can be seen that the optimal size of the word-based HMM is 24 states per model. Since the small amount of data available, the number of Gaussians in the HMM was chosen to be one Gaussian component per state. The baseline in the three cases of matched-model non-pathological speech, mismatched-model non-pathological-pathlogical speech and matched-model pathological speech is shown in Table 4. The results over the four train-test sets present a WER mean of 30% for the mismatched condition and 20% for the matched condition, showing the difficulty of the task of ASR for pathological speech.

Training set	Test set	Mean WER	STDV WER
Non-path.	Non-path.	6.91%	2.33%
Non-path.	Path.	29.97%	2.94%
Path.	Path.	19.15%	2.50%

Table 4: *Baseline results.*

The results of the frequency and time MATE are shown in Tables 5 and 6. In both cases there is a noticeable reduction in the WER. Frequency MATE reduces the WER in the mismatched condition a 11.22% and up to 17.04% for the matched condition. Time MATE reaches a 8.12% reduction of WER in the mismatched training, while improving a 11.19% in the matched condition. These results demonstrate that the impact of the time and frequency variability on the ASR performance as it has been studied in this work can be reduced by this or other kind of more complex models in which this sources of variability are taken into account.

6. Conclusions

In this work, we have presented the results of a recent Vocal Tract Length Normalization algorithm applied to the task of an ASR system in the presence of speech uttered by people with several kinds of speech disorders. We have shown how speech pathologies affect the speech signal and distort its time and frequency characterictics. The baseline results obtained with our corpus show clearly the bad influence of these variations in pathological speech in an ASR task. The final recognition results show an improvement up to 17% in the WER. These results point out that a warping method like MATE can reduce fruitfully time and frequency variations in speech for ASR when taken separately. But for future works, it should be considered the possiblity of reducing the impact of time and frequency distortions with a mixed time-frequency warping method. Better results should be expected by taking into account interaction between both domains.

Training set	Test set	Mean WER	STDV WER	Improvement
Non-path.	Non-path.	5.37%	1.41%	22.22%
Non-path.	Path.	26.61%	2.32%	11.22%
Path.	Path.	15.89%	2.28%	17.04%

Table 5: *Frequency MATE results.*

Training set	Test set	Mean WER	STDV WER	Improvement
Non-path.	Non-path.	4.94%	2.00%	28.56%
Non-path.	Path.	27.54%	2.58%	8.12%
Path.	Path.	17.01%	2.03%	11.19%

Table 6: *Time MATE results.*

7. References

[1] Darley F.L., Aronson A.E., Brown J.R., "Differential diagnostic patterns of dysarthria", Journal of Speech and Hearing Research, 12(2):246–269, 1969.

[2] Croot K., "An acoustic analysis of vowel production across tasks in a case of non fluent progressive aphasia", In Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP), 1998.

[3] Green P., Carmichael J., Hatzis A., Enderby P., Hawley M., Parker M., "Automatic speech recognition with sparse training data for dysarthric speakers", In Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech), 1999.

[4] Navarro-Mesa J.L., Quintana-Morales P., Pérez-Castellano I., Espinosa-Yañez J., "Oral Corpus of the project HACRO (Help Tool for the Confidence of Oral Utterances)", Department of Signal and Communications, University of Las Palmas de Gran Canaria, May 2005.

[5] Monfort M., Juárez-Sánchez A., "Registro Fonológico Inducido (Trajetas Gráficas)", Ed. Cepe, Madrid, 1989.

[6] Lee L., Rose R., "A frequency warping approach to speaker normalization", IEEE Transactions on Speech and Audio Processing, 1(6):49–60, 1998.

[7] Miguel A., Lleida E., Rose R., Buera L., Ortega A., "Augmented state space acoustic decoding for modeling vocal variability in speech", In Proceedings of the 9th European Conference on Speech Communication and Technology (Eurospeech), 2005.

[8] Pitz M., Ney H., "Vocal tract normalization equals linear transformation in cepstral space", IEEE Transactions on Speech and Audio Processing, 13(5):930–944, 2005.

[9] Deller J.R., Hsu D., Ferrier L.J., "On the use of Hidden Markov Modelling for recognition of dysarthric speech", Computer Methods and Programs in Biomedicine, 35:125–139, 1991.