PRONUNCIATION VARIATION AND ITS IMPORTANCE IN SPEECH RECOGNITION

MANUEL PARRAS RUIZ DE AZÚA 850604-A592
JULIA RAMÍREZ GARCÍA 881218-A207

PRONUNCIATION VARIATION AND ITS IMPORTANCE IN SPEECH RECOGNITION

ABSTRACT

In this report we have studied the pronunciation variation and its importance in speech recognition. For it, we have focused in the level of modeling, and, being more precise, in the lexicon level, acoustic and language models. Besides, an useful algorithm is analyzed: the rule learning algorithm. In pronunciation variability, we study the differences between phonemes and syllables and which is better for being recognized.

1. INTRODUCTION

When anyone hears something about automatic speech recognition (ASR), he, or she, thinks about commands given by voice to a machine. In certain way, it is true. ASR can be used in as many ways as we can imagine, although they are not simple.

Behind ASR, there are a lot of techniques and algorithms to make the recognition in an universal way, without the need of changing them depending on the gender, the age, or other features of the speaker; just a program capable of recognizing different characteristics of voices in an effective way.

We have to consider that automatic speech recognition (ASR) would be very easy if all the words were pronounced in an unique way. However, this is true for some reasons and words are almost always pronounced in different ways. We are going to discuss the most important sources of pronunciation variation.

It is in this moment when the pronunciation has to be taken into account for making a good speech recognition program. Differences in pronunciation are difficult to be treated when talking about ASR. Dealing with different accents or ways of pronouncing, the trickiest part is the one related with training, because we have to upgrade the initially simple pronunciation models by learning pronunciation rules based on algorithms rather than pronunciation variants from the data. [1]

One of the most known phenomenons when we talk about phonology is the existence of variations in the words pronunciation. These variations can be produced by crossword articulations, regional effects and other mechanisms, which can be speaker-dependent, as well as others that are very general and are considered as speaker-independent.

When the speech becomes less formal, the syllabic structure of words differs from the normal one: speech rate usually increases and there may be some changes in pitch and loudness. There is also free variations in which the speaker feels free to choose from among different pronunciations of the same word.

Another important influence in speech is the interlocutor, since it is known that people speak in different way depending on the person they are talking to.

If we speak in a strict way, we could say that almost all ASR research nowadays is about modelling pronunciation variation.

Taking into account most of the modern ASR systems, most of them are based in subword accoustic models (phone components, phones, phonemes, syllables, …). In order to describe how the entries in the vocabulary that we want to recognize can be pronounced, a lexicon is needed.

However, to improve the lexicon by introducing pronunciation variants can also increase the similarity between words and therefore produce confusion that would lead to a bad recognition accuracy. Because of this, we have to avoid adding any possible pronunciation variation to the lexicon. We think that the chance of success is more likely when a limitation on the acceptance of pronunciation variants is considered.

Hence, the introduction of pronunciation variants in the training is expected to lead into a benefit, provided that this happens in a controlled way.

The departure point would be a basic lexicon that contains a single pronunciation for each word. This lexicon is going to be the *reference lexicon* and the pronunciations contained in it will be the *reference pronunciations*. We consider that most of the pronunciation variants of the word can be reached by only applying *pronunciation rules* to the reference pronunciations.

During the training we would impose a hierarchy on the rules and we would also add some exception rules which do not produce pronunciation variants but do affect the production of such variants by other rules.

Although using isolated words makes an ASR system to work better and in an easier way, it certainly does not do the same to the speaker, because making a pause between every single word is very unnatural.

“There is a solution between the ease of modeling at the lexicon level and the need to introduce the model cross-word variation.” [2] This solution is called multi-words and consists on a sequence of words that are treated as one entity in the lexicon, and the variations that result when the words are strunged together are modeled by including different variants of these multi-words. However, it is important to note that “with this approach only a small portion of cross-word variation is modeled.” [2]

Another thing that we have to take into account that distinguishes the different approaches to modeling pronunciation variation in ASR is the source from which information on pronunciation variation is derived. A distinction can be made between data-driven and knowledge-based methods: “the mayor difference between these two types of approaches is that in the former case the assumption is that the information on pronunciation variation has to be obtained in the first place. In knowledge-based approaches, on the other hand, it is assumed that this information is already available in the literature.” [2]

This justifies a *data-driven* scheme relying on *already trained acoustical models* to determine pronunciation rules from orthografically transcribed speech data.

"In knowledge-based studies, information on pronunciation variation is primarily derived from sources that are already available." [2] This sources can be linguistic studies in pronunciation variation and pronunciation dictionaries. "However, these sources do not provide enough information, they only provide the form of the possible variants, while quantitative information on the frequency of these alternative variants has to be obtained from the acoustic signals, as it is the case for data-driven methods." [2]

1.1. RULE LEARNING ALGORITHM [1]

We are going to present "a method for building stochastic pronunciation models incorporating acceptable pronunciation variants" [1]. The method is based on a rigid pronunciation rule formalism that is expanded with a set of rules and imposes a number of constraints presented as negative rules or exception rules. "The obtained pronunciation networks can replace the single pronunciation word models commonly used in speech recognizers" [1].

The goal of this algorithm is to learn, from relevant orthographically transcribed speech data, a consistent set of pronunciation rules describing *frequently* occurring pronunciation modifications (compared to the *reference pronunciation* of the lexicon) with a *minimal context*. The general outline of the algorithm is the following:

1.- *Transcription generation*: For each utterance in the training data set, two phonemic transcriptions are determined: a reference transcription T_{ref} that describes how it should have been pronounced according to the reference lexicon, and an expert transcription T_{ex} describing how the phrase was pronounced according to a non-human expert.

2.- *Alignment of transcriptions*: Line up T_{ex} with T_{ref} and derive from it the correspondences between the phonemes of the reference and those of the expert transcription.

3.- *Candidate rule generation*: Analyze the correspondences between the reference and the expert transcription so that regions are identified in the reference transcription that looks different from corresponding regions in the expert transcription.

4.- *Rule pruning*: Organize the rules in a hierarchical way.

5.- *Negative rule identification*: Use the application likelihoods computed in the previous step to decide on the positive or negative nature of each rule.

2. LEVEL OF MODELING

To model the variation, three levels of modeling are set because of the three components of the recognition engines of most ASR systems: the lexicon, the acoustic models and the language model. Three models are needed to obtain a good recognition system, from the top to the bottom. [2]

2.1. LEXICON

At this level, the modeling is carried out by adding pronunciation variants to the lexicon. This is because the speech recognizer has more chances to select a transcription belonging to the correct word if multiple transcriptions of the same word are given. Lower error rates should be reached with this method; however, the acoustic confusability in the lexicon may increase ("the added variants can be confused with those of other entries in the lexicon"), adding new errors that did not exist before. Therefore, an appropriate selection of the pronunciation variants must be done. Besides, the use of multi-words to model cross-word variation is also a good idea. [2]

2.2. ACOUSTIC MODELS

At the level of acoustic models, it is recommended to use forced recognition. New transcriptions of the signals are computed this way, so that they are used to train new acoustic models and, then, this new acoustic models, used to do forced recognition in a loop way. This process is called *iterative transcribing*. [2]

2.3. LANGUAGES MODELS

We suppose that we have a speech input signal, X, that must be recognized by finding the string of words, W, that maximizes P(X|W)*P(W). P(W) can be computed with the help of N-grams. As said before, the most common way to model pronunciation variation at the lexicon level is to add pronunciation variants to the lexicon. To deal with these pronunciation variants at this level three methods are proposed:

- Method 1: This is the easiest. It consists of adding the variants to the lexicon, so that the LMs are not changed. This method just applies the probability of a variant to belong to a word, not the variant probability itself. This leads to a sub-optimal solution.
- Method 2: This method uses the variants and their probabilities to compute the N-grams. For it, it needs a transcribed corpus that contains information about the realized pronunciation variants in order to maximize the above equation.
- Method 3: In this method, an intermediate level is introduced: P(X|V)*P(V|W)*P(W). The goal now is to find the string of words, W, and the string of variants, V, that maximizes this equation. [2]

3. PRONUNCIATION VARIABILITY

The syllable is preferred over some other form of multi-phone representation for ASR because no all words are spoken in canonical form. In fact, observing the pattern of pronunciation variation in spontaneous speech, it is shown that it is far from egalitarian. Words differ a lot in terms of their frequency of occurrence. The Switchboard lexicon illustrates this magnitude: “The most frequent words occur far more frequently than the least (Fig. 1). The ten most common words account for approximately 25% of all lexical occurrences”. Analysing these most frequently occurring words, it is demonstrated that the biggest amount belong to the group of pronouns, articles, conjunctions and modal or auxiliary verbs.

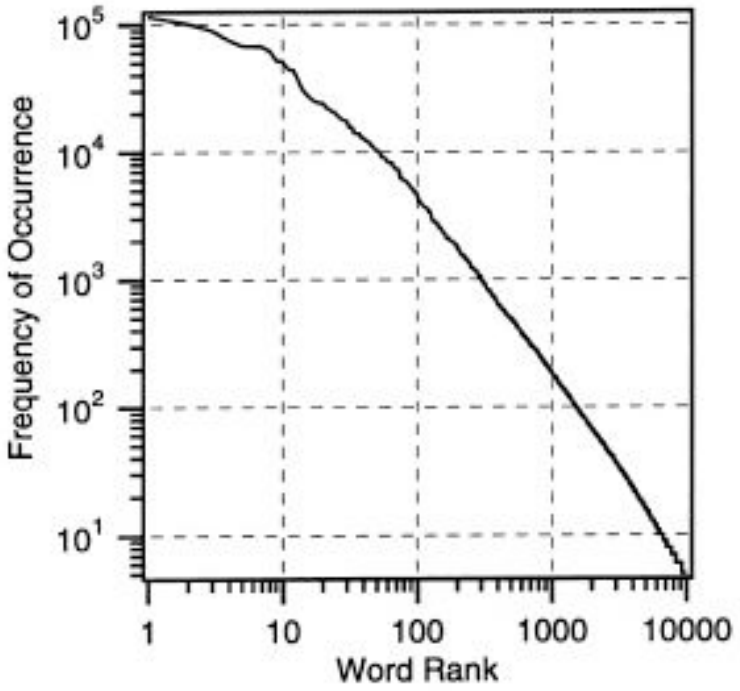

Figure 1.- “The frequency of occurrence for the 10,000 most frequent words in the Switchboard corpus, organized in rank order of frequency. Total number of distinct words in the corpus is 25,923.” [3]

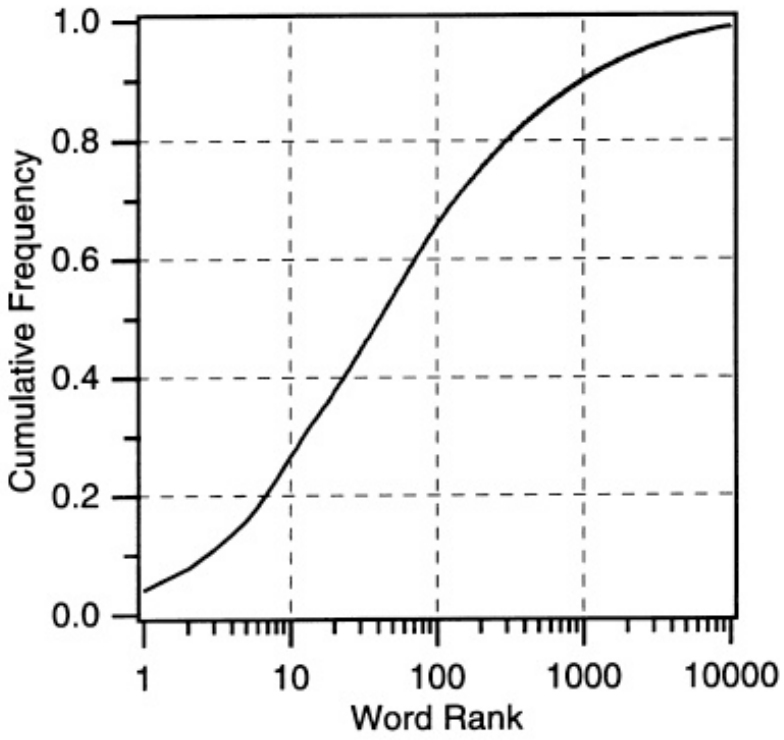

Figure 2 .- "Cumulative frequency of occurrence as a function of word frequency rank for the 10,000 most frequent lexical items in the Switchboard corpus." [3]

From the statement said above and, as it can be seen in Table 1, only 22% of the Switchboard lexicon is composed of monosyllabic forms, approximately 80% of the words are just one syllable in length. In spontaneous speech, it is weird to use lexicon consisting of three or more syllables (less than 5%). This data study is very useful in order to achieve a good decoding of the speech stream. Of course, this data only concerns to English language, although it can be extended to others.

# Syllables	Usage (%) (All)	Usage (%) (STP)	Lexicon (%)
1	81.04	78.42	22.39
2	14.30	16.31	39.76
3	3.50	3.72	24.26
4	0.96	0.95	9.91
5	0.18	0.23	3.21
6	0.02	0.03	0.40

Table 1.- "The proportion of words consisting of n-syllables for the entire Switchboard for the entire corpus, the portion of the corpus, the portion of the corpus phonetically transcribed and lexicon." [3]

Nouns or adjectives (words of high information valence) tend to be pronounced in a canonical fashion, whereas pronouns, conjunctions and articles (common lexical items) are pronounced in a very personal

and individual way. This suggests that the information valence associated with words and syllables can be very important for the design of ASR systems, as the regular lexicon just contains canonical pronunciation for each input. Besides, the most frequent words tend to be spoken faster because of their inherent predictability.

Table 2 shows the pronunciation variability for the 100 most common words in the phonetically segmented portion of the Switchboard Transcription. [3]

Table 2 .- Pronunciation variability for the 100 most common words in the phonetically segmented portion of the Switchboard Transcription Corpus[a]

	Word	*N*	#Pr.	Most common pronunciation	%Tot
1	I	649	53	ay	53
2	and	521	87	ae n	16
3	the	475	76	dh ax	27
4	you	406	68	y ix	20
5	that	328	117	dh ae	11
6	a	319	28	ax	64
7	to	288	66	tcl t uw	14
8	know	249	34	n ow	56
9	of	242	44	ax v	21
10	it	240	49	ih	22
11	yeah	203	48	y ae	43
12	in	178	22	ih n	45
13	they	152	28	dh ey	60
14	do	131	30	dcl d uw	54
15	so	130	14	s ow	74
16	but	123	45	bcl b ah tcl t	12
17	is	120	24	ih z	50
18	like	119	19	l ay kcl k	46
19	have	116	22	hh ae v	54
20	was	111	24	w ah z	23
21	we	108	13	w iy	83
22	it's	101	14	ih tcl s	20
23	just	101	34	jh ix s	17
24	on	98	18	aa n	49
25	or	94	23	er	36
26	not	92	24	m aa q	24
27	think	92	23	th ih ng kcl k	32
28	for	87	19	f er	46
29	well	84	49	w eh l	23
30	what	82	40	w ah dx	14
31	about	77	46	ax bcl b aw	12
32	all	74	27	ao l	24
33	that's	74	19	dh he s	16
34	oh	74	17	ow	61
35	really	71	25	r ih l iy	45
36	one	69	8	w ah n	78
37	are	68	19	er	42
38	I'm	67	9	q aa m	26
39	right	61	21	r ay	28
40	uh	60	16	ah	41
41	them	60	18	ax m	23
42	at	59	36	ae dx	8
43	there	58	28	dh eh r	22
44	my	58	9	m ay	66
45	mean	56	10	m iy n	58
46	don't	56	21	dx ow	14
47	no	55	8	n ow	77
48	with	55	20	w ih th	35
49	if	55	18	ih f	41
50	when	54	18	w eh n	31
51	can	54	28	kcl k ae n	15

[3]

Table 2 (Continued)

	Word	*N*	#Pr.	Most common pronunciation	%Tot
52	then	51	19	dh eh n	38
53	be	50	11	bcl b iy	76
54	as	49	16	ae z	18
55	out	47	19	ae dx	22
56	kind	47	17	kcl k ax nx	21
57	because	46	31	kcl k ax z	15
58	people	45	21	pcl p iy pcl l el	44
59	go	45	5	gcl g ow	83
60	got	45	32	gcl g aa	15
61	this	44	11	dh ih s	47
62	some	43	4	s ah m	48
63	would	41	16	w ih dcl	29
64	things	41	15	th ih ng z	52
65	now	39	11	n aw	69
66	lot	39	9	l aa dx	47
67	had	39	19	hh ae dcl	24
68	how	39	11	hh aw	53
69	good	38	13	gcl g uh dcl	27
70	get	38	20	gcl g eh dx	13
71	see	37	6	s iy	80
72	from	36	10	f r ah m	28
73	he	36	7	iy	39
74	me	35	5	m iy	87
75	don't	35	21	dx ow	14
76	their	33	19	dh eh r	25
77	more	32	11	m ao r	56
78	it's	31	14	ih tcl s	20
79	that's	31	20	dh eh s	16
80	too	31	6	tcl t uw	60
81	okay	31	17	ow kcl k ey	45
82	very	30	11	v eh r iy	36
83	up	30	11	ah pcl p	34
84	been	30	11	bcl b ih n	51
85	guess	29	8	gcl g eh s	42
86	time	29	8	tcl t ay m	62
87	going	29	21	gcl g ow ih ng	13
88	into	28	20	ih n tcl t uw	14
89	those	27	12	dh ow z	42
90	here	27	11	hh iy er	25
91	did	27	13	dcl d ih dx	23
92	work	25	8	w er kcl k	66
93	other	25	14	ah dh er	26
94	an	25	12	ax n	28
95	I've	25	7	ay v	46
96	thing	24	9	th ih ng	52
97	even	24	7	iy v ix n	40
98	our	23	9	aa r	33
99	any	23	11	ix n iy	23
100	we're	23	8	w ey r	25

4. CONCLUSIONS

To develop reliable acoustic models so that a robust speech recognition is achieved, phonetic characterization of spoken language is needed. These recognition systems must face important difficulties, such as variability in speaking style or the acoustic background. Newest ASR have just realized that the syllable may be a more basic organizational unit than the phone at the acoustic level. Therefore, at the lexicon level, using syllables instead of the more traditional phonemic sequences is more likely to obtain a better recognition of the speech signal.

As said before, by adding pronunciation variants to the lexicon, pronunciations variation is modeled. This method improves recognition performance. However, certain words ("that", for example) have numerous variants with many different frequencies of occurrence. Thus, the confusability at the lexicon may increase (so that the recognition success may decrease) if many variants of a large number of words are included. Therefore, a variant selection must be performed. An obvious criterion is the frequency of occurrence, so that frequent variants produce better recognition than infrequent variants. In any case, the right solutions have not been found yet and more research should be done but in which direction? At least, one important axiom has changed: before, a fixed assumption was that speech was made up of discrete segments, which were phonemes, nowadays syllables are considered, but the idea that speech can be phonologically represented as a sequence of discrete entities has proved to be untenable.

5. REFERENCES

[1] Cremelie, N., Martens, J-P., 1999. *In search of better pronunciation models for speech recognition*. ELIS, University of Gent, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

[2] Strik, H., Cucchiarini, C. 1999. *Modeling pronunciation variation for ASR: A survey of the literature*. A^2RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

[3] Greenberg, S. 1999. *Speaking in shorthand - A syllable-centric perspective for undestanding pronunciation variation*. International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, USA

[4] Greenberg, S. 1997. *On the origins of speech intelligibility in the real word.* In: Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels. Pont-a-Mousson, France, pp. 23-32