Do speech recognizers prefer female speakers?

Martine Adda-Decker and Lori Lamel

Spoken Language Processing Group
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{madda,lamel}@limsi.fr

ABSTRACT

In this contribution we examine large speech corpora of prepared broadcast and spontaneous telephone speech in American English and in French. Starting with the question whether ASR systems behave differently on male and female speech, we then try to find evidence on acoustic-phonetic, lexical and idiomatic levels to explain the observed differences. Recognition results have been analysed on 3-7h of speech in each language and speech type condition (totaling 20 hours). Results consistently show a lower word error rate on female speech ranging from 0.7 to 7% depending on the condition. An analysis of automatically produced pronunciations in speech training corpora (totaling 4000 hours of speech) revealed that female speakers tend to stick more consistently to standard pronunciations than male speakers. Concerning speech disfluencies, male speakers show larger proportions of filled pauses and repetitions, as compared to females.

1. INTRODUCTION

In the early days of automatic speech recognition, female speech was widely considered as more difficult to automatically recognize than male speech. In fact many early results were only reported for male speakers. However, progress made during the 1980s/1990s in automatic speech recognition, has led to high performance transcription systems for found speech such as broadcast news (BN) and quite reasonable performance for conversational telephone speech (CTS). At the same time, gender is not longer considered a major issue.

To model gender-specificities in automatic speech recognition, it is common to estimate gender-dependent acoustic models. Techniques such as Speaker Adaptive Training (SAT) [1] and unsupervised adaptation (MLLR) [9] also reduce the influence of speaker and gender. In addition vocal tract length normalization (VTLN) [2] processing on the acoustic parameter space allows adjustment for the speaker's physical characteristics. Whereas the proportions of speech from male speakers is dominant in BN data, the inverse tendency can be observed for CTS corpora (about 40% male).

In the next section, we examine the overall recognition results (obtained with gender-specific acoustic models) for male and female speakers on a total of more than 20 hours of transcribed speech in French and English for broadcast news and for conversational telephone speech. In the following two sections, we make use of the available transcribed training data (over 3500 hours) to try to relate the differences in recognition results with aspects such as lexical usage, pronunciation and disfluency differences between male and female speakers.

2. GENDER-DEPENDENT ASR RESULTS

As a starting point, we examine recently obtained recognition results in English and French using BN and CTS [10, 4, 8, 7] in an attempt to investigate and explain gender performance differences. For English the data correspond to DARPA RT04 BN and CTS development test sets. The French BN data are the *Technolangue*-ESTER [3]. development data and the French CTS set comes from LIMSI internal resources. Whereas published word error rates generally give averaged male and female error rates, in the following we report separate figures (word error which sums the different types of errors: substitutions, deletions and insertions) according to gender.

Broadcast news speech

Automatic transcription of general broadcast news in French and English yields word error rates which are close to 10%. Table 1 gives error rates for male and female speakers. For English the average word error rate for the female speakers is 2.6% lower than the male speakers, whereas in French the error rates are much closer (0.7% lower for females). Part of this difference may be due to the higher proportion of interviewees that are male, and the higher number of interviews in the English data. The interviewees tend to speak in a less prepared manner than the anchors and reporters, and often the acoustic quality is lower. This argument may give at least a partial explanation for lower recognition rates with male speech. The smaller gender difference observed for French may be because all of the speech comes from the radio, whereas the English data is mostly from TV. Previously BN results have typically reported higher word error rates on TV than on radio shows.

	%Wer	%corr	%sub	%del	%ins
Fr-m (5h)	9.4	91.7	5.6	2.7	1.1
Fr-f (2h)	8.7	92.4	5.5	2.1	1.1
En-m (3h)	12.0	89.5	7.1	3.4	1.5
En-f (1h)	9.4	92.4	5.4	2.3	1.8

Table 1: Recognition results on English and French broadcast news. For English BN dev04 data have been used. For French the data corresponds to the ESTER dev set.

Conversational telephone speech

Table 2 summarizes the word error rates by gender for the CTS data. The English conversational speech data come from the Switchboard and Fisher corpora distributed by LDC. The callers typically do not know each other, are supposed to speak about an assigned topic, and are paid for their participation. The French conversations were often carried out between friends and/or family members, and therefore have a very casual speak-

ing style, which may partially account for the high word error rate compared to English. Two other main factors for the high error rate on French is that the pronunciation dictionary has not been sufficiently adapted to spontaneous speech variants, and the available data for acoustic and language modeling is an order of magnitude below those available in English.

	%Wer	%corr	%sub	%del	%ins
Fr-m (1.5h)	45.2	57.8	27.6	14.5	3.0
Fr-f (4.5h)	37.9	65.6	23.5	10.9	3.5
En-m (1.5h)	15.7	86.3	8.4	5.2	2.0
En-f (1.5h)	13.7	88.4	7.9	3.7	2.0

Table 2: Recognition results on English and French conversational telephone corpus. For English H5, the dev04 data have been used. For French the data come from LIMSI internal resources.

As observed for the broadcast news data, female speakers have lower error rates on the conversational data. Even if there is no longer an asymmetry in roles between male and female speakers on telephone conversations, as was suggested for BN speech, more data are available for females here. This is always a convincing argument for better results. Looking more in detail into different error types, we can observe that the deletion rate is significantly higher for male speakers than for females, and the insertion rates tend to remain similar. This tendency can reflect either a faster speaking rate for males, or incomplete pronunciations. In section 4 we will address this question in more detail.

3. LANGUAGE USAGE: LEXICAL DISTRIBUTION

Further investigations are conducted on the training data which amount to 3500 hours of transcribed speech over all conditions. Table 3 shows the amount of transcribed speech in English (3020 hours) and French (455 hours) used for the subsequent analyses for both the BN and CTS data.

French	male		female	
	# words	duration	# words	duration
BN	2703 k	270 h	906 k	90 h
CTS	236 k	25 h	709 k	70 h
English	male		female	
BN	4262 k	420 h	2942 k	300 h
CTS	10956 k	1000 h	13666 k	1300 h

Table 3: Corpus sizes for BN and CTS, English and French.

If different error rates are observed for male and female speech on BN data, the reason may be due to different proportions of speech types related to different roles played by the speakers depending on the gender. In particular, we know that a majority of interviewed personalities are males. In this situation speech is less prepared and less predictable than, for example, news headlines. To get some insight for this hypothesis we looked at words that are used significantly more in male or in female speech. To do so, for each word its proportion in the gender-specific corpus is computed and the words, for which the proportion varied by more than a fixed percentage across gender, are extracted. This kind of analysis only focuses on the top N most frequent words (here N=1000). Significant differences for English BN speech can be observed on the top 1000 words in training data. There are 50 words which occur more often (than 0.1%) in male speech and they are almost exclusively function words. With the same criterion 70 words are found in female speech, and a larger proportion of content words are found: *CNN, weather, headline, rain, police, snow, showers, president, temperatures....* A large proportion of the female speech seems to involve headlines and reports, whereas the male speech also includes less formal portions. This argues in favor of better results for female speech.

If the error rate differences are to be explained by a less careful articulation for male speech than for female speech, we can examine the frequency of occurrence of words for which different lexical entries are available depending on standard or sloppy speaking styles. Among the more frequent candidates here are *yes*, *yeah* for English, and equivalent items *oui* and *ouais* for French. Table 4 shows both the frequency of occurrence as well as the overall importance in male and female speech respectively. In Table 4 other lexical entries for agreement *okay* and the backchannel *uhhuh*, *right* are added.

CTS data				
French	male		female	
oui (std)	2829	1.2	14111	2.0
ouais (slop)	8197	3.5	14799	2.1
total	11026	4.7	28910	4.1
English	male		female	
yes (std)	10424	0.1	20802	0.2
yeah (slop)	206308	1.9	229212	1.7
uhhuh (bck)	45680	0.4	89548	0.7
right (bck)	66557	0.6	87353	0.6
okay (bck)	27139	0.2	37033	0.3
total	262412	3.3	339562	3.4

Table 4: Agreement and backchannel words in English and French conversational speech over telephone for.standard and sloppy speaking styles.

For both French and English the more sloppy pronunciation of *yes* is preferred by a large amount. Nonetheless when examining percentages of standard pronunciations (*yes, oui*), female speakers tend to produce twice as many as male speakers. There is no difference in the use of the backchannel word *right*, however the familiar item *uhhuh* is twice as frequent for female speakers than for male. Although *right* can also occur as a word or a response, when we looked at the usage in the CTS data we found that many of the occurrences were backchannel or discourse markers.

The type and frequency of disfluencies is potentially interesting to find differences in speech usage with respect to gender. In the following we will focus on filled pauses and repetitions. Filled pauses are transcribed in detailed manual audio transcripts or automatically aligned for other transcripts. Speech repetitions can be roughly estimated by the count of occurrences of identical word sequences in trigram language models. Table 5 shows the corresponding counts and percentages. The overall proportion filled pauses produced in spontaneous speech depends on the communication situation, the degree of familiarity between protagonists, the urgency of the information content, the emotional load. A priori these latter factors are the same for both genders since the majority of conversations are between speakers who do not know each other and discuss an assigned subject. For the given communication setups under study here, male speakers produce 50% more filled pauses than female speakers in both languages. The situation is similar for repetitions. These observations are consistent with those made in [11]

CTS data				
French	male		female	
filled pauses	10595	4.5	20321	2.9
repetitions	7600	3.2	17499	2.5
English	male		female	
filled pauses	319223	2.9	272035	2.0
repetitions	207504	1.9	183711	1.3
BN data				
French	male		female	
filled pauses	2169	0.08	623	0.07
repetitions	12236	0.45	2516	0.3
English	male		female	
filled pauses	31976	0.75	7559	0.27
repetitions	12853	0.30	5169	0.18

Table 5: Disfluencies in CTS and BN data for French and English. Filler words and single word repetitions are reported.

4. SPEECH : PRODUCTION, ARTICULATION, PRONUNCIATION AND ACOUSTICS

If fewer ASR transcription errors are consistently observed with female speech, the explanation is probably not a simple acoustic one. Female speech is considered more difficult to process, at least for pitch extraction and formant measurements. Female voices have a higher fundamental frequency than male voices, and typically have shorter vocal tracts and therefore higher formant frequencies so the useful bandwidth is smaller. Whereas for telephone speech, lower word error rates could be explained by a better fit of the female voice ambitus to the telephone bandwidth, lower error rates are consistently observed on large-band processed speech.

Can lower error rates then be related to a more careful articulation or a more canonical-like pronunciation in female speech? This hypothesis might be supported by the traditional role of female speakers in oral communication, and particularly by their role in language acquisition in parent-child relations and in education. Do male speakers more easily adopt a sloppy speaking style whereas women stick to more standard pronunciations?

From an automatic processing perspective, if the uttered words fit the foreseen pronunciations used for acoustic word modeling, each of the word's phone models can be aligned with a corresponding portion of the acoustic signal. Hence each aligned phone should reach a minimum duration of at least 50ms (corresponding to a 20 phone/second rate). If the uttered words are significantly shortened with respect to allowed pronunciations, then the rate of minimum duration segments (0.03s) should increase.

To find a partial answer to this question, we analysed the speech corpora after carrying out a forced alignment using context-independent phone models so as to allow for a higher use of pronunciation variants. Pronunciation reductions, in the sense of incomplete or shorter pronunciations, can be hypothesized on schwa vowels ("a", "amount", "I agree"), on long words ("California", "administrative", "necessarily", "probably", "particularly"), on sequences of function words ("did you", "you have", "I am not"), on discourse markers ("you know"), on complex consonant clusters at word boundaries ("east coast", "best friend"), on idiomatic expressions ("a little bit", "a lot of", "make fun of them").

To explore data with the idea of shortened pronunciations in mind, we are interested in segment durations. As the transcribed audio data have been segmented into phones, the proportion of minimum duration phones in male/female speech can be measured. This can be considered as a good indicator of reduced pronunciations. Figure 1 shows the phone length distribution of the Fisher/Switchboard training corpus (totaling more than 50M phones). For the CTS-English about 15% of the phone segments have a minimum duration of 0.03 seconds. For male and female speech this proportion is around 16.5% and 13.5% respectively. Whereas the overall curves are very similar for durations above 0.04s, there is a significant difference for the shortest segment length.

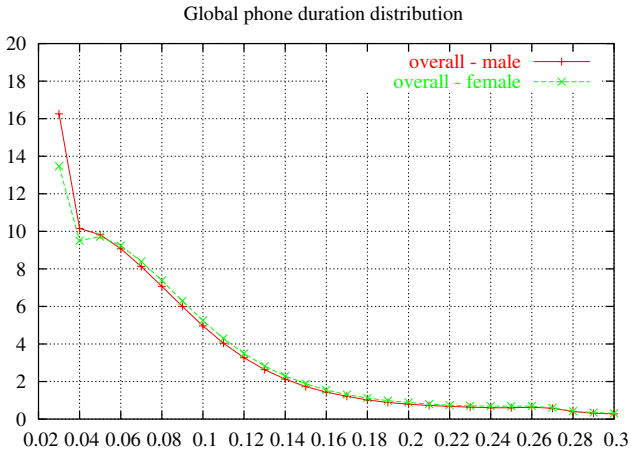

Figure 1: Distribution of phone length for male and female speech in the English CTS training data (Switchboard/Fisher, 53M phone segments). About 16.5% of the male phones and 13.5% of the female phones have the minimum duration of 0.03s.

% minimum duration		
consonants	male	female
t	31	27
d	26	23
v	22	17
n	20	17
h	20	15
dh	20	16
ng	19	13
syllabic-n	18	14
r	17	14
vowels	male	female
schwa	36	30
I	28	23
U	24	20
u	24	20
$\wedge$	19	16
ϵ	19	17
r-schwa	18	13

Table 6: CTS English corpus. Phones with a high percentage of minimum duration segments.

We next look more in detail which phonemes contribute most to this minimum segment length rate. Table 6 shows vowels and

consonants which have the highest rates of minimum duration segments (0.03s) in CTS English. Whereas it is interesting to compare these figures across speaking styles and languages, we simply focus here on gender differences. The trend of minimum duration can be observed on the same phonemes in the same order, for both male and female. However we always observe a lower percentage for female speakers by roughly 3%, which supports what is seen in Figure 1.

/m/ phone segments	male		female	
subset	#occ.	avg. dur.	#occ.	avg. dur.
all	807k	70ms	928k	70ms
money	9k	90ms	9k	90ms
minimum	8k	40ms	9k	50ms

Table 7: CTS English (Fisher) data. Number of /m/ phone segments (*1000) and average duration in male and female speech. Statistics are given in general and for two carrier words *money*, *minimum*.

In order to look into this more closely, we selected two high frequency words (*money, minimum*) for both male and female speakers in the CTS data. Both words occur about 9k times for each gender. The word initial /m/ in *money* has an average duration 90ms for both genders. The intra-word /m/ in *minimum* has an average duration about 50ms for females and 40ms for males. (The overall average duration for /m/ across all contexts is about 70ms for both genders.) More information can be gleaned for the phone duration distribution shown in Figure 2, where it can be seen that the phone duration for the word-initial, stressed-syllable /m/ in *money* has a nice bell shape. The reduced (middle) /m/ in *minimum* has a bimodal distribution for both male and female speakers, but there is a larger peak for a minimal duration for the males.

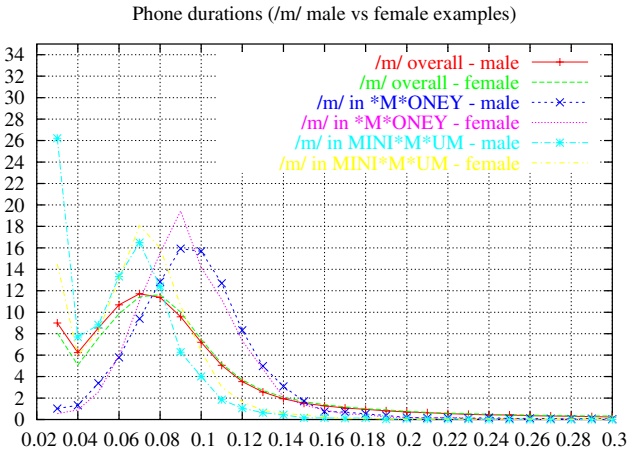

Figure 2: Comparison of /m/ phone durations (given in percentage) for male and female speech. There are roughly 1 million /m/ segments for each gender in the overall curve and 9k for the two example carrier words *money* and *minimum*.

5. DISCUSSION AND PERSPECTIVES

In this contribution we have carried out a first analysis of the differences in speech recognition performance with respect to gender. This study has looked at two types of data broadcast news and spontaneous conversational telephone speech in English and French. For all conditions female speakers had better average recognition results than males. Absolute differences amount to 0.7% for French BN, 2.6% for English BN and 2% for CTS English, where absolute error rates range from 10% (BN) to 15% (CTS English). For CTS French the difference between female and male speakers is particularly high (7.3%) since the absolute error rate is quite high on this corpus.

These studies open a range of questions as to why there is this performance difference [6]. As mentioned above there are likely to be multiple factors that come into play. For the more careful speech found in news broadcasts, speakers are usually selected for a certain speaking quality, thus the observed differences may be due to a larger proportion of non-professional male speech in the data. The confirmation of this trend for conversational speech supports the idea that some of the difference may be attributed to the traditional role of women in language acquisition and education. Independent of the reasons for the gender differences, this study has given us ideas for improving the acoustic models and recognition lexicon to account for the observed duration changes.

REFERENCES

[1] T. Anastasakos and J. McDonough and R. Schwartz and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," *ICSLP'96*, 2:1137-1140.

[2] A. Andreoum T. Kamm and J. Cohen, "Experiments in Vocal Tract Normalisation," *CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[3] Galliano et al., "The Ester Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News", submitted to Interspeech 2005.

[4] J.L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, H. Schwenk, "Where Are We In Transcribing BN French? " submitted to Interspeech 2005.

[5] J.L. Gauvain, C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussain Mixture Observation of Markov Chains," *IEEE Trans. on SAP*, **2**(2), pp. 291-298, April 1994.

[6] W. Labov, The intersection of sex and social class in the course of linguistic change. Language Variation and Change 2:205-254, 1990.

[7] L. Lamel, J.L. Gauvain, G. Adda, M. Adda-Decker, L. Canseco, L. Chen, O. Galibert, A. Messaoudi and H. Schwenk, Speech Transcription in Multiple Languages, *Proc. IEEE ICASSP'04*, **III**:757–760, 2004.

[8] J.L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, F. Lefèvre, "Conversational Telephone Speech Recognition," *IEEE ICASSP'03*, Hong Kong, April 4-5, 2003.

[9] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, **9**(2):171-185, 1995.

[10] R. Schwartz, T. Colthurst, N. Duta, H. Gish, R. Iyer, C.-L. Kai, D. Liu, O. Kimball, J. Ma, J. Makhoul, S. Matsoukas, L. Nguyen, M. Noamany, R. Prasad, B. Xiang, D-X. Xu, J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda et L. Chen, "Speech Recognition in Multiple Languages and Domains: the 2003 BBN/LIMSI EARS System," *IEEE ICASSP'04*, **III**:753–756, 2004.

[11] E.E. Shriberg, Preliminaries to a Theory of Speech Disfluencies, PhD thesis, University of California at Berkeley, 1994. PH.D. Thesis.