Modelling Human Speech Perception in Noise

A. Kabir, M. Giurgiu

***Abstract*-**Human auditory system of speech perception tries to find out by applying computational technique how human perceive speech. The difference between the current state of art automatic speech recognition (ASR) and human speech perception (HSP) is the prior knowledge about a given speaker such as speaking style, gestures, eye movements and so on. Therefore if an ASR is feed by the knowledge of a given speaker, then it could be said as HSP system. This paper presents the preliminary research in order to develop a HSP system in Romanian with a view to make it language independent. Acoustic analysis and speech glimpsing are investigated in order to do so. The principal findings are machine tends to recognize noisy speech with a more or less constant recognition rate, but still with a poor recognition rate in compare to their human counterparts, and acoustic parameters have less influence in recognizing noisy speech. In addition, a Romanian speech corpus which we named as RO-GRID is collected in ordered to use as the common material in speech perception and automatic speech recognition. Utterances are simple, syntactically identical phrases such as "muta bronz cu p 2 agale." The corpus is annotated at the phoneme, syllable and word level and is available on the website for research use.

***Keywords*—**Romanian Speech Corpus, Hidden Markov Models, Speech Intelligibility, Speaker Intelligibility, Vocal Tract Length Normalization, Glimpsing Speech.

I. Introduction

HUMAN listeners are much better than automatic speech recognizer (ASR) at recognizing speech in everyday noise conditions [1]. This scientific challenge led to two approaches of modeling speech perception: macroscopic models and microscopic models. Macroscopic models provide an indication of overall speech intelligibility in masking and reverberation where microscopic models apply automatic speech recognition technique in order to provide listeners response to individual tokens [1].

It is pointed in [1] that the results of microscopic modeling is assuring but it requires very large amount of training data. GRID corpus collected by the University of Sheffield, United Kingdom is an excellent solution to these problems in order to support joint computational-behavioral studies in speech perception. But when it comes to carry out perceptual experiments in another language rather than English, GRID corpus is unable to answer this call as it is an English speech corpus. Moreover English language has sufficient numbers of speech corpus but Romanian language lacks of suitable speech corpus for the purpose of research. It is one out of many reasons to collect a large multi-talker speech corpus in Romanian language.

The corpus is seriously influenced by GRID corpus which has sentences such as bin blue at f 2 now of the form <command: 4> <color: 4> <preposition: 4> <letter: 25> <digit: 10> <adverb: 4> [1, 3]. The new corpus which we named RO-GRID uses only 10 very highly confusable letters reducing the numbers of sentences from 1000 to 400 per speaker in compare to GRID corpus.

There is another reason exists why we are motivated to collect a Romanian corpus similar to GRID corpus. The reason is carry out cross language study. Many works have been done using the GRID corpus and the results are readily available. It would be interesting to see if these hypotheses also hold in another language like Romanian. If these hypotheses hold, then language dependency might not be a problem anymore for carrying out certain experiments.

The rest of the paper is organized as follows: section II will provide the details about the corpus design and its collection, section III will depict the post processing after the collection of corpus, section IV will present experimental results to show the intelligibility of the collected speech materials as well as to show the performance of humans which will be used as the reference for the evaluation with the performance of machines, section V will evaluate the performance of acoustic models in clean environment, section VI will describe the influence of acoustic parameters in clean and noisy environments, section VII will provide glimpse analysis to show the performance of machine in noisy conditions and finally we will provide discussions which will be followed by a conclusion.

II. The RO-GRID Corpus

A. Sentence Design

Each sentence of the RO-GRID corpus has designed by a sequence of six words illustrated in the Table I. Of the six words, color, letter, and digit were designated as "keywords". In the letter position, 10 highly confusing letters were used because ASR is normally able to recognize non confusing words. Eight of them are consonants and two of them are vowels. Each speaker uttered all combinations of the three keywords, producing to a total of 400 sentences per speaker. The rest of the three words of a sentence, command,

Manuscript received March 18, 2011. This work was supported by the the EC-funded Project Marie Curie Research Training Network MRTN-CT-2006-035561 S2S ("Sound to Sense", www.sound2sense.eu).

A. Kabir, is with the Department of Telecommunications, Technical University of Cluj-Napoca, 26 Baritiu Street, 400027 Cluj-Napoca, Romania. (phone: +40-264-401807; e-mail: ahsanul.kabir@com.utcluj.ro).

M. Giurgiu, is with the Department of Telecommunications, Technical University of Cluj-Napoca, 26 Baritiu Street, 400027 Cluj-Napoca, Romania. (e-mail: mircea.giurgiu@com.utcluj.ro).

preposition, and adverb were designated as "fillers." Each filler has four alternatives and randomly chosen while designing a sentence. Filler words provide some variation in contexts for the neighboring keywords.

B. Speakers

Nine male and three female speakers aged 20-28 years contributed to the corpus. Speakers are undergraduate students and PhD students at the Technical University of Cluj-Napoca (UTCN) and originated from various regions of Romania. Every speaker speaks Romanian as their first language. All but one speaker is bilingual who has a family connection with Hungary. He had born in Romania, had spent his life in Romania but also speak Hungarian. The mean age of the population is 25 years.

C. Speech Collection

Recordings were made in a reasonably quiet TV studio located inside UTCN. The recording software implemented jointly in MS Visual Basic 2008 (Express Edition) and Microsoft SQL Server 2008 (Express Edition). Speech materials were collected from a desk mounted high quality beyerdynamic opus 89 professional microphone. The microphone was connected to the Yamaha MW12c USB mixer device which was connected to a computer via USB interface. Collection of speech material was under computer control.

TABLE I. STRUCTURE OF RO-GRID CORPUS

Command	Color	Prep	Letter	Digit	Adverb
vezi	negru	La	p, t, d, g, j,	0-9	putin
(look)	(black)	(at)	b, v, h, o, u		(few)
muta	verde	De			agale
(move)	(green)	(by)			(slowly)
pune	bronz	In			acolo
(put)	(bronze)	(in)			(there)
sari	auriu	Cu			afara
(jump)	(golden)	(with)			(outside)

Sentences were presented on a computer screen placed in front of the participants and had 5 seconds to speak every sentence. Participants were advised to speak in a natural manner as if they are used to communicate with others. They were asked to speak adequately quickly to fit into the 5 seconds time window. Participants could repeat the sentence if they felt it necessary as if they made a mistake during production or if part of the spoken utterance did not fit into the designed 5 seconds window. The recording supervisor also had control of the recorded speech and could accept or reject recordings, or initiate re-recordings or move on to the next sentence.

Recordings were divided into five recording sessions for each participant including the repeat session. A total of 100 sentences made up each of the first four recording session and the repeat session was made up depending on the number of mistake made by the individual participant so that were necessary to repeat. Participants were asked to take an optional break if the feel it necessary. Recordings were completed by more or less 50 minutes for each participant.

III. POST PROCESSING

A. Voice Activity Detection

Each recording session for a particular speaker was collected as long sound file where .wav was the extension of speech file. The long sound file was processed frame by frame in order to compute energy and zero crossing rates of each frame. Then the long sound file was broke up into constituent uttered sentences based on analysis of energy and zero crossing rates. In a word, the algorithm is able to detect silence in between two consecutive sentences. Renaming convention of the sentences was to take the first letter of the first five uttered words and the second letter of the sixth word. It provides a unique filename for each uttered sentence for an individual speaker. Therefore if the uttered sentence is muta verde la h 2 agale, then mvlh2g.wav would be the name of the corresponding sound file.

B. Corpus Transcription

The next task was to transcribe the corpus. To provide a manual phonetic transcription for such a large speech corpus would require ample amount of efforts and time. To avoid these difficulties, scientifically established technique is to perform forced alignment by using hidden markov models (HMM). For this purpose, the corpus has been divided into two sets where 5% of the corpus belongs to the train set and the rest belongs to the test set. Train set covers every word, syllable and phoneme of the entire corpus and has been transcribed manually in the word, syllable, and phoneme level. Then initial models have been generated by training the train set using hidden markov model toolkit (HTK). After that, train set has been bootstrapped from these models and trained in speaker independent (SI) as well as speaker dependent manner (SD). Finally, forced alignment was performed by using SD models in order provide transcription of the test set. Finally the automatic transcription was also converted into .TextGrid extension which can be imported into PRAAT for further processing. Figure 1 shows the transcription which has been imported into PRAAT.

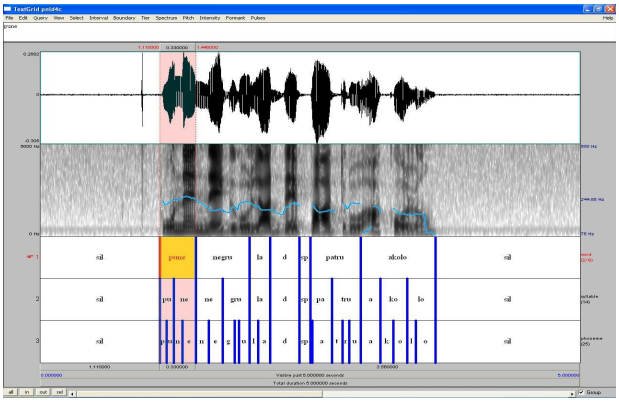

Figure 1. Transcription of the sound file *pnld4c.wav* imported into PRAAT

IV. Intelligibility Test

A. Listeners

Four male and one female listener aged 23-27 years contributed in this pilot study. Most of them are PhD students at the Technical University of Cluj-Napoca (UTCN) and originated from various regions of Romania. Every speaker speaks Romanian as their first language and none of them is reported to have hearing problems. The mean age of the population is 25 years.

B. Speech and Noise Materials

Forty sentences have been drawn from the RO-GRID corpus in a random manner. Then the initial and trailing silence has been removed from each sentences and added to the 8-talker babble noise at 6, 4, 2, 0, -2, -4, -6 dB of signal to noise ratio (SNR). Therefore, a total of 1600 sentences (1400 noisy and 200 clean) were used in the preliminary intelligibility test.

C. Procedures

Listeners were presented 1400 noisy sentences and 200 clean sentences and the task were to identify color, letter, and digit and to enter their results through software. Figure 2 shows the intelligibility test software. Listener is able select the type of speech by controlling testing session, play a sound, and enter the result. Listener is not able to proceed to the next sentence until entering the results. The software is particularly useful because it provides fast and accurate data entry. Most listeners were able to complete the test within 35 minutes.

Figure 2. User interface of intelligibility test software for fast and accurate data entry

D. Experimental Results

Listener's answers were interpreted by correctly identified keywords (colors, letters, and digits). Figure 3 shows identification of keywords as a function SNR across the entire population. Colors are identified more accurately followed by digits and letters respectively. The difference with letter increased as SNR decreased. It could be noted that when SNR is -6dB, digits are identified more accurately than colors. It can be also noted that the number of choice available at each keywords influences the results. Though letters and digits have ten choices each, digits have been identified more accurately than the letters because the letters used in this corpus are highly confusing letters.

Figure 4 shows the overall intelligibility of female speakers and male speakers in percentage. Keywords from female speakers are identified more accurately than the male speakers and the difference increased substantially when SNR became more and more negative. Therefore gender has a significant effect on speaker intelligibility.

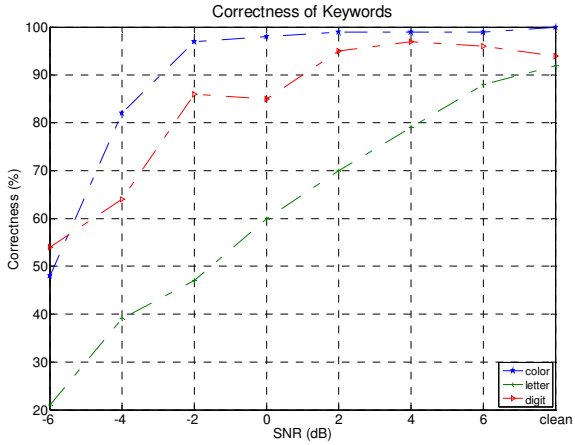

Figure 3. Correctly recognized keywords

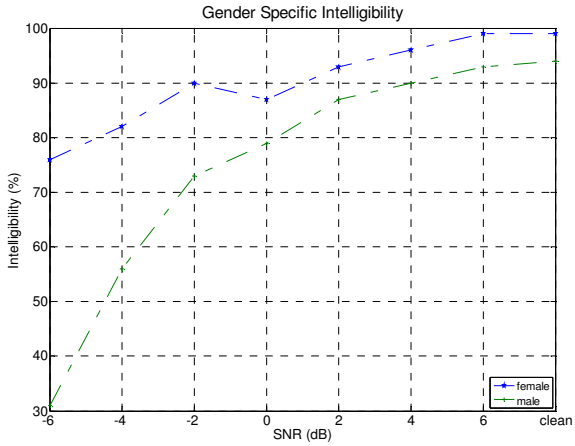

Figure 4. Intelligibility of male and female speakers as a function of signal to noise ratio (SNR)

Now 7 noisy conditions of this perceptual test have been classified into three groups. They are low SNR (-6, -4 dB), medium SNR (-2, 0, 2 dB) and high SNR (4, 6 dB). Figure 5 shows intelligibility of each keyword in low SNR, medium SNR and high SNR. The color *auriu*, the letter *h*, and the digit *7* have been identified more accurately in comparison to other colors, letters, and digits in very noisy condition. On the other hand, every color, letter, and digit is identified more or less accurately in less noisy condition.

Figure 6 shows intelligibility of each speaker in low SNR, medium SNR and high SNR. Second speaker has higher intelligibility in very noisy condition where every speaker has

more or less high intelligibility in less noisy condition. Not surprisingly, second speaker is a female speaker.

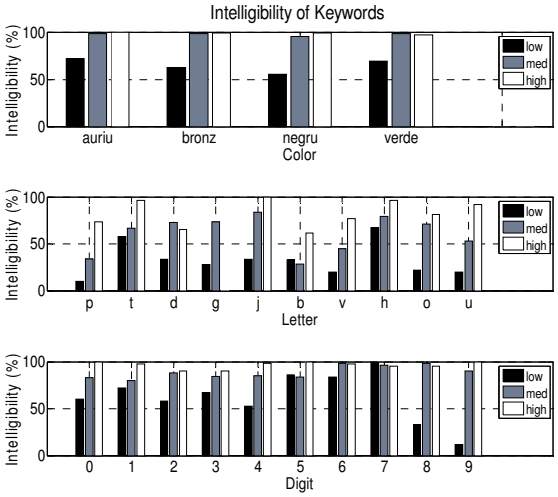

Figure 5. Intelligibility of individual keywords

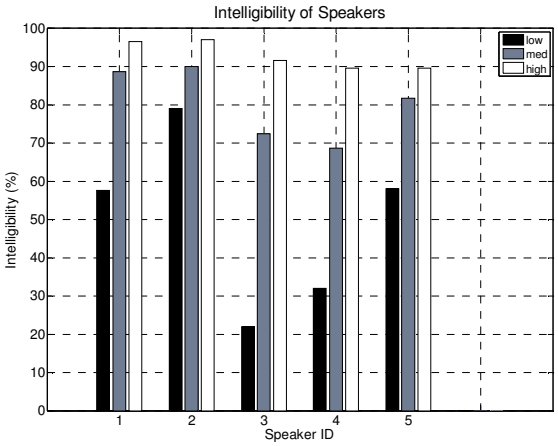

Figure 6. Intelligibility of each speaker

V. Evaluation of Acoustic Models

An HMM based automatic speech recognizer is used to evaluate the acoustic models. Phone recognition rate is 85.79% when testing is carried out in a speaker independent manner and with the insertion of the optional short pause among the words of sentences. But recognition rate increased substantially when testing is carried out in a speaker dependent manner. Best performance is achieved for speaker of id2 and speaker of id3 with a recognition rate of 95.78% and 95.59% respectively where worst performance is noticed for speaker of id9 and speaker of id1 with a recognition rate of 91.45% and 91.99%. Figure 7 shows the phone recognition rate when testing is carried out in speaker dependent and gender dependent manner with the insertion of optional short pause. Though the difference is not significant, but it can be noted that male speakers are recognized more clearly than the female speakers. Male speakers have an overall phone recognition rate of 94.25% where female speakers have an overall phone recognition rate of 93.96. Figure 8 shows the phone recognition rate for each speakers. It necessarily signifies a very good set of acoustic phone models with this recognition rate.

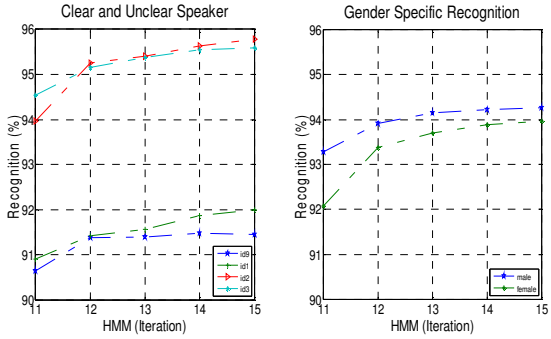

Figure 7. Speaker dependent phone recogniton and gender dependent phone recogniton

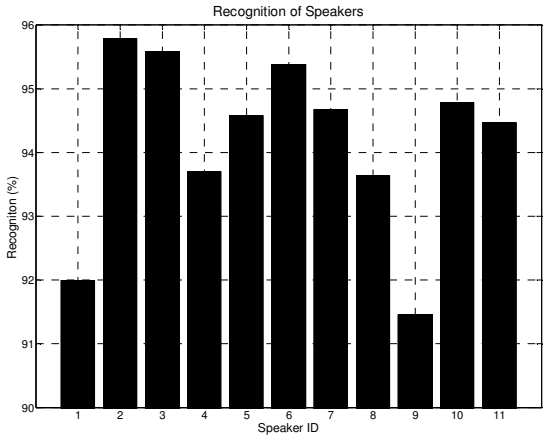

Figure 8. Phone recognition rate of each speaker in clean speech

VI. Acoustic Measurements

A series of acoustic measurements was registered in order to look into the reasons for the intelligibility of the RO-GRID sentences in clean and noisy environments.

A. Vowel Duration

Duration was computed from transcription of each utterance and can also be represented as the speech rate because all sentences of RO-GRID are of the identical length. Mean vowel duration (ms) computed over the entire population varies from 64.54 ms to 227.15 ms and shown in Figure 9.

B. Formant Estimation

First three vowel formant frequencies (F1, F2, and F3) are estimated by classical liner predictive coding (LPC), Burg

Algorithm implemented in PRAAT, and Auto-Regressive method for male speakers and female speakers at the frame which is located at the mid-point of the time interval corresponding to each vowel instance of the RO-GRID corpus. Formant estimation technique is described in [7].

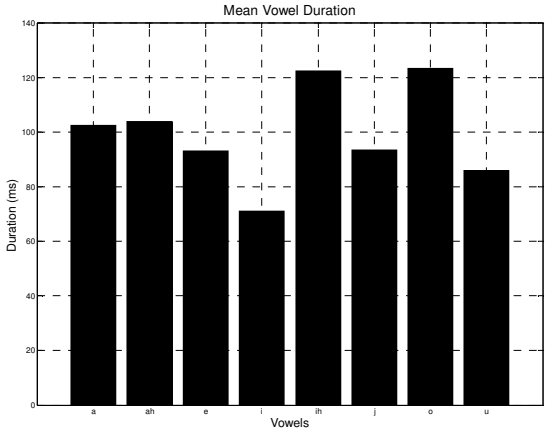

Figure 9. Average vowel duration of RO-GRID corpus

The average values for the first three formants estimated by three techniques are shown in Table II, Table III, Table IV. Vowel quadrilateral diagram given by the formants estimated by three techniques is shown in Figure 10, Figure 11 and Figure 12.

TABLE II. AVERAGE F1, F2, AND F3 BY LPC (CLEAN SPEECH)

Vowel	Male			Female		
	F1	*F2*	*F3*	*F1*	*F2*	*F3*
/a/	606	1174	2253	562	1161	2062
/ah/	505	1205	2206	526	1153	1969
/e/	423	1423	2370	425	1078	2291
/i/	337	1356	2387	357	1256	2465
/ih/	350	1248	2286	380	1236	2184
/j/	339	1465	2365	381	1370	2426
/o/	429	922	2143	452	915	2057
/u/	332	953	2099	365	971	2100

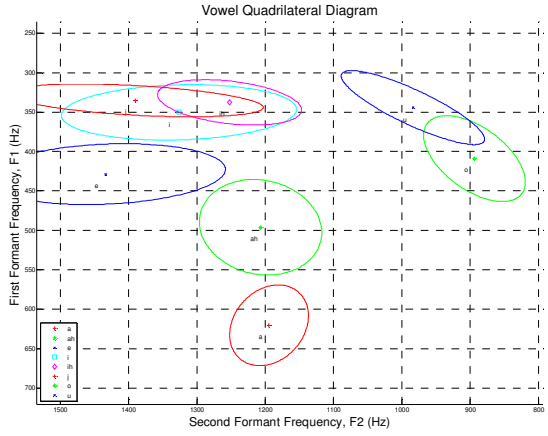

Figure 10. Vowel space given by the formants estimated through the LPC algorithmburg

TABLE III. AVERAGE F1, F2, AND F3 BY PRAAT (CLEAN SPEECH)

Vowel	Male			Female		
	F1	*F2*	*F3*	*F1*	*F2*	*F3*
/a/	700	1270	2500	756	1408	2656
/ah/	637	1368	2671	728	1457	2924
/e/	470	1827	2657	531	1973	2912
/i/	574	2045	2776	602	2240	2940
/ih/	401	1316	2655	465	1448	2768
/j/	353	2077	2734	391	2354	2914
/o/	525	1057	2704	570	1094	2901
/u/	414	1178	2685	437	1227	2818

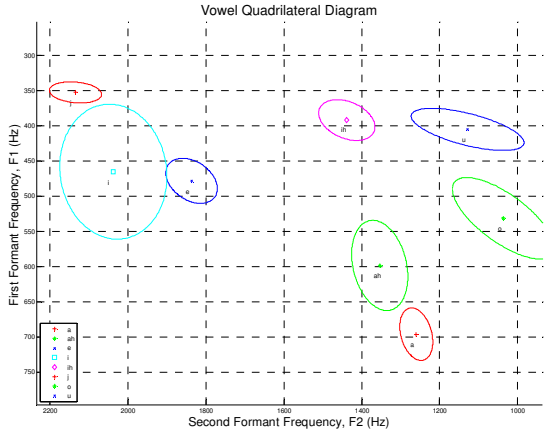

Figure 11. Vowel space given by the formants estimated though the burg algorithm in PRAAT

TABLE IV. AVERAGE F1, F2, AND F3 BY AUTO-REGRESSIVE METHOD (CLEAN SPEECH)

Vowel	Male			Female		
	F1	*F2*	*F3*	*F1*	*F2*	*F3*
/a/	584	1136	2143	528	1117	1977
/ah/	484	1114	2078	500	1078	1765
/e/	417	1456	2398	431	1237	2409
/i/	324	1379	2382	330	1150	2473
/ih/	348	1080	2303	366	1086	2073
/j/	318	1525	2477	353	1733	2666
/o/	432	861	2267	450	914	2127
/u/	336	809	2243	381	964	2315

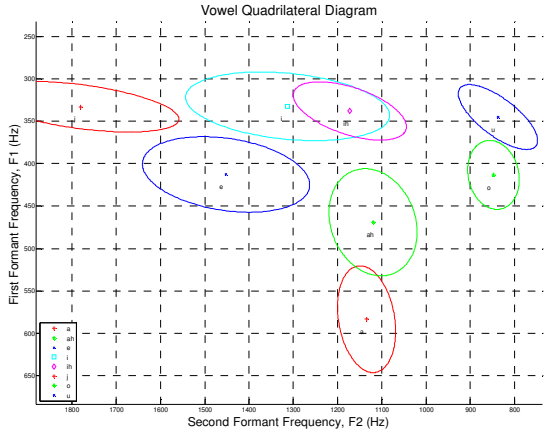

Figure 12. Vowel space given by the formants estimated through Auto-Regressive Method

C. *Vocal Tract Length Normalization*

Formants provide an opportunity to apply feature based vocal tract length normalization (VTLN) which is computationally economic in compare to model based VTLN. Normalization factor has been computed according to the fixed-formant pattern model which is described in [8].

For each vowel independently, average formant frequencies for F1, F2 and F3 are computed for the entire RO-GRID corpus in order to provide the reference point for formant frequency warping. A normalization factor is applied to each formant which minimizes the distance between formant frequencies of an individual speaker and the reference speaker. The average of the normalization factor computed over all vowels is taken as the normalization factor of the specific speaker.

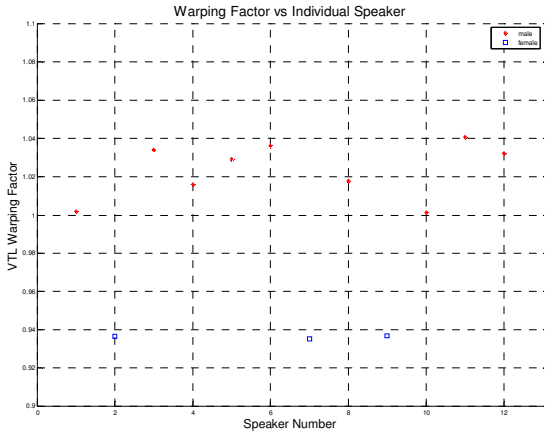

Figure 13. Warping factor for each speaker from the formants estimated by the burg algorithm in PRAAT

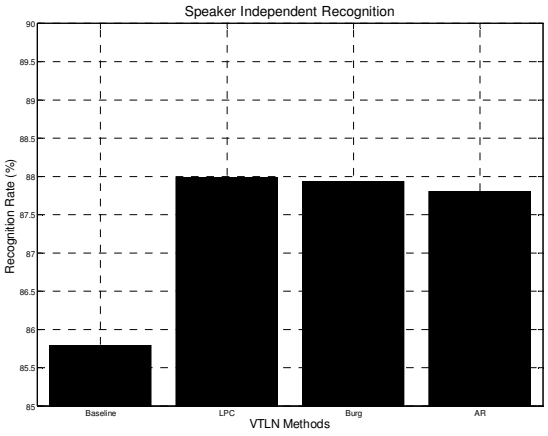

Figure 14. Recognition rate of clean speech after applying VTLN

Figure 13 shows normalization factor of each speaker of the RO-GRID corpus plotted against speaker number. Separation between the male speakers and female speakers could be easily identified. Notably, the reference speaker has a warping factor of one.

Figure 14 shows the performance of the baseline system (without normalization), and after applying VTLN where warping factor is estimated by formant frequencies computed by LPC, Burg, and AR method. SD recognition rate is improved by maximum 2.2% when VTLN is applied. In addition, there is no substantial difference exists in recognition rate by the computation of formants from different methods.

D. *Vowel Normalization*

A scatter distribution of formant frequencies is the usual result when both type of adult male speakers and adult female speakers are present in the dataset. It is because the vocal tract length is significantly different between adult male speakers and adult female speakers. But a compact distribution of formant frequencies would be the result if the vocal tract shapes are normalized to a reference speaker. It is proved from Figure 15, Figure 16, and Figure 17 that formant frequency distribution became compact after applying normalization technique.

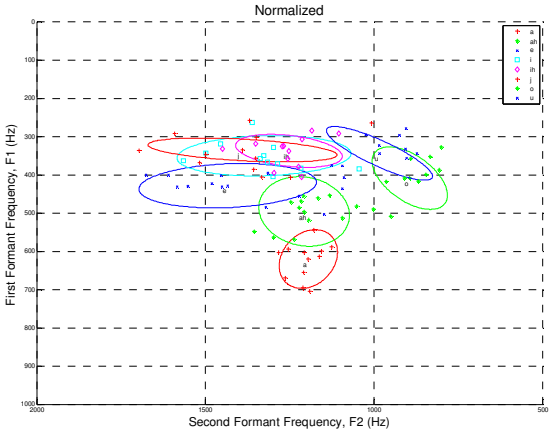

Figure 15. Normalized vowel space given by LPC method after applying VTLN.

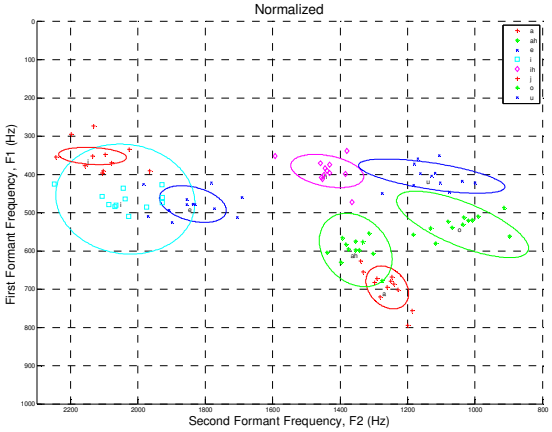

Figure 16. Normalized vowel space given by the Burg algorithm implemented in PRAAT after applying VTLN.

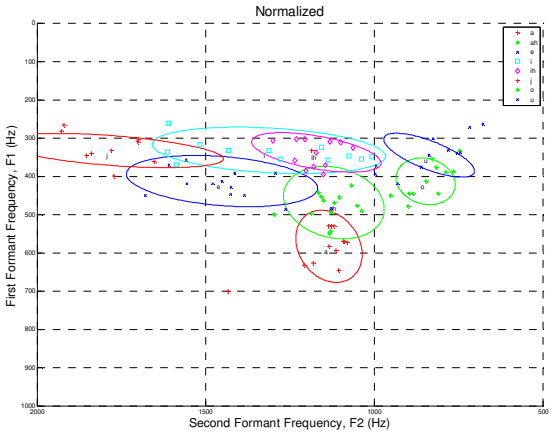

Figure 17. Normalized vowel space given by LPC method after applying VTLN.

VII. Glimpse Analysis

Speech glimpses are defined by the difference of auditory excitation patterns between speech and noise where speech is with energy concentrated in compact regions of the spectro-temporal plane [3]. According to [3:14], "Even at highly unfavorable SNRs, there will be local regions where the speech stands clear of the noise floor. The size, shape and spectro-temporal position of these glimpses will be dependent on the characteristics of the speaker. For example, a speaker with a peakier long term spectrum is likely to produce more glimpses than a speaker with a flatter spectrum. Availability of reliable speech glimpses is likely to be a contributing factor to the intelligibility of the speech."

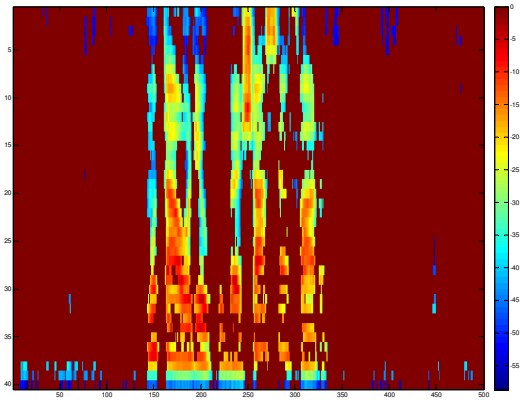

Figure 18. Glimpsing area for the utterance *muta airiu cu b 6 afara* with 8-talker babble noise at 6dB SNR and having 3dB of threshold.

Therefore, a glimpsing model which engages passing the speech signal through a bank of 40 gammatone filters with centre frequencies ranging from 50Hz to 7500Hz linearly spaced in the ERB scale, was applied to each speaker of the RO-GRID corpus. Within each channel, the Hilbert envelope is calculated and a leaky integrator of 8ms was imposed. This procedure also repeats for 8-talker babble noise which was used to produce noisy speech for intelligibility tests. Then a comparison between speech signal and noise signal leads to the computation of local SNR. Then noisy utterances are labeled as reliable speech if the local SNR at a given point is greater than the threshold. In our case, it is 3dB. Figure 18 shows speech glimpse where the utterance is mixed with 8-talker babble noise at 6dB SNR and having 3dB of threshold.

Then speech is resynthesized after applying glimpsing technique by gammatone filterbank. Resynthesized speech is the representation of useful glimpses, possibly occurring at different times and occupying different regions of the spectrum, and assumes that listeners somehow integrate those glimpses to hear out the target speech from the noisy environments.

Finally resynthesized speech is passed to the automatic speech recognizer which is trained by clean speech and after applying VTLN. Recognition rate is less than 20% when recognition is performed by the baseline models. Recognition rate even drops down further by 2-3% when recognition is performed by the VTLN models. It necessarily means that VTL information is lost or somehow damaged during the resynthesis process.

VIII. Discussion

This study tries to integrate the way how human perceives speech into automatic speech recognition. Recognition rate by humans drops as the speech becomes noisier. Moreover performance in ASR improves when VTLN is applied. Surprisingly, performance decreases for the noisy speech when VTLN was applied after the speech resynthesis. Therefore, it signifies that VTLN could be a positive technique for clean speech but needs to be reconsidered for noisy speech. In addition, it is seen by the perceptual experiments that female speakers are more intelligible than the male speakers even in noisy conditions which could be exploited into ASR in noise to model better human speech perception system.

IX. Conclusions

Research about Romanian language is suffering severely due to lack of suitable speech corpora. RO-GRID corpus has been collected as a part of data collection in order to carry out perceptual tests and automatic speech recognition in Romanian. Pilot intelligibility test implies that speech materials are quite good enough to identify in quiet and low noise conditions. An HMM based ASR also proved that collected speech materials can used for very good acoustic modeling. Please visit www.sound2sense.eu for the complete corpus and various kinds of transcriptions (phone, syllable, word).

It is also seen that humans are excelling by far than the machine while recognizing noisy speech. But for complex keywords like letters are difficult to recognize by human where machine can be better in this case. Moreover, machine tends to

recognize noisy speech with a more or less constant recognition rate, but still with a poor recognition rate. Enhanced signal processing techniques or noisy estimation techniques could be a potential solution to this problem.

ACKNOWLEDGMENT

This work has been carried out in the frame of the EC-funded Project Marie Curie Research Training Network MRTN-CT-2006-035561 S2S ("Sound to Sense", www.sound2sense.eu).

REFERENCES

[1] M. Cooke, J. Barker, S. Cunningham and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition." Journal of the Acoustical Society of America, vol. 120, 2006.

[2] M. Cooke, O. Scharenborg, "The Interspeech 2008 Consonant Challenge." in Proc. of Interspeech, 2008, pp. 1-4.

[3] J. Barker and M. Cooke, "Modeling Speaker Intelligibilty in Noise." Speech Communications, vol. 49, 2007, pp. 402-417.

[4] I. Amdal, J. Svendsen, "FonDat1: A Speech Syntehsis Corpus for Norwegian," in Proc. of LREC, 2006.

[5] I. Amdal, O.M. Strand, J. Almberg, and T. Svendsen, "RUNDKAST: An Annotated Norwegian Broadcast News Speech Corpus," in Proc. of LREC, 2008.

[6] G. Barrese, M.F. Bontempi, P.J. Wundes, P.E. Boim, "Development of a digital hearing aid with advanced processing algorithms," in *Proc. of The 8th WSEAS International Conference on Signal, Speech, and Image Processing,* 2008, pp. 67-71.

[7] A. Kabir, J. Barker, & M. Giurgiu. "Robust Formant Estimation: Increasing the Reliability by Comparison among Three Methods," in *Proc. of The International Conference on Circuits, Systems, Signals,* 2010, pp. 341-344.

[8] A. Kabir, J. Barker, & M. Giurgiu. "An Approach to Vocal Tract Length Normalization by Robust Formants," in *Proc. of The International Conference on Circuits, Systems, Signals,* 2010, pp. 345-348.

[9] F. Martinex-Licona, O. Munoz-Texzocotetla, A. Martinez-Licona, J. Goddard, "Analysis of emotions in Mexican Spanish speech," in *Proc. of The 8th WSEAS International Conference on Signal, Speech, and Image Processing,* 2008, pp. 67-71.

[10] A. Jorschick, "Sound to Sense Corpora," Internal S2S Technical Report, 2009.

Ahsanul Kabir was born in Bagerhat, Bangladesh in 1982. He received B.Sc. in Computer Science & Engineering from Khulna University of Engineering & Technology (KUET), Bangladesh in 2006 and also received M.Sc. in Wireless Communications Systems Engineering from the University of Greenwich, United Kingdom in 2007. He awarded Marie Curie Research Fellowship in 2008 and currently working as a researcher at the Technical University of Cluj-Napoca, Romania. His research interests includes digital signal processing, speech signal processing, and free space optical communications.

Professor Mircea Giurgiu received M.Sc. in Electronics and Telecommunications Engineering from the Technical University of Cluj-Napoca, Romania in 1990, and Ph.D. in Telecommunications from the same university in 1996. He became a full professor at the Department of Telecommunications since 2006. His research interests include engineering methods for automatic speech recognition, text-to-speech synthesis, digital content management, and web-based applications for education.