Review on Multilingual Blended Speech Recognition using Gaussian Mixture Model for Non-Dictionary Words

Seema Kumari[1], Gaurav Garg[2],
[1]*Perusing M-Tech, Department of CSE, AITM at Palwal, Haryana, India*
[2]*Assistant Professor, Department of CSE, AITM at Palwal, Haryana, India*
(E-mail: seemakumarip2@gmail.com)

***Abstract*—**Utilization of multilingual blended language in everyday spoken model is getting to be normal and is acknowledged as being grammatically right. Anyway, machine acknowledgment of blended language spoken speech models are a test to a traditional Multilingual Blended Speech Recognition. There are things about the best way to empower the acknowledgment of Multilingual Blended Speech Recognition. Toward one side of the spectra is to utilize acoustic models of the total verbal communication with the set of the blended language to empower recognition while on the opposite end of the spectra is to utilize a language distinguishing proof module pursued by language-dependent speech acknowledgment to do the acknowledgment. Every one of this has its own ramifications. In this paper, we approach the problem of blended language discourse acknowledgment by utilizing accessible assets and demonstrate that by reasonably developing a proper articulation lexicon and changing the language model to utilize blended language, one can accomplish a decent acknowledgment precision of spoken blended language. Therefore using N-Gram and Gaussian Mixture Model for Multilingual Blended Speech Recognition system will be developed with more accuracy and effectiveness.

Keywords*—Speech Recognition, Acoustic Model, Automatic Speech Recognition, Hidden Markov Model, Gaussian Mixture Model.*

I. INTRODUCTION

Automatic speech recognition, in the future alluded as ASR, changes over spoken words into content. In the previous decade, numerous calculations had been considered and created to improve the execution of ASR frameworks. Mainstream utilizations of ASR, for example, voice seek, voice control and spoken exchange framework, and so on., had additionally been generally utilized. An ASR framework by and large incorporates two noteworthy segments: the front-end and the decoder. As appeared in Figure 2.1, the front-end separates highlight perceptions O from the information discourse flag S, in order to acquire an appropriate representation of discourse. While the decoder uses the predefined acoustic model, language model and lexicon to recoup words W from the element perceptions O. However consider a call centre in a metropolitan city which needs to take into account individuals talking distinctive dialects based on multilingual blended language. This requires every one of the operators in the call centre to have the capacity to convey in numerous dialects or speech patterns which are in all respects questionable. A conceivable arrangement can be to learn the language of the caller and teller, in light of the language, direct the caller to an operator who can banter in that language expertly. In a comparative vein, in a discourse empowered application, having recognized the language of the guest, a language-explicit discourse acknowledgment mechanism can be utilized to take into account the guest. Obviously, this sort of framework can't work when individuals utilize blended language discourse, regardless of whether one knew the blend of dialects being used, in light of the fact that the language move is visit. As of late there has been expanded enthusiasm for blended language acknowledgment, anyway the work has been confined to a blend of hindi and English (amalgamation). Blended language model acknowledgment is in its beginning phases of research and to the best of our insight, there is no work detailed in the writing for India explicit language blend. There are two noteworthy particular systems to manufacture blended language programmed discourse acknowledgment (Mixed Language- Automatic Speech Recognition), to be specific multi pass and one pass structure. In a multi-pass ML-ASR (Mixed Language-Automatic Speech Recognition), the accurate occurrence in the verbally expressed discourse where language switch happens is resolved and the language of the discourse recognized. When the language of the discourse section is known, relating language subordinate programmed discourse acknowledgment (ASR - Automatic Speech Recognition) is utilized to perceive the discourse fragment. Note that a normal ASR is language explicit and utilizes acoustic model (AM), language model (LM) and an articulation vocabulary (pronunciation lexicon) worked for that language to perceive spoken discourse. The AM. LM and PL (pronunciation lexicon are built from language-explicit discourse and content corpus through a preparation procedure. In the one pass approach, an ASR is manufactured (in particular, AM, LM, and PL) which envelops both the dialects in the blended language. This empowers ML-ASR multilingual blended language communication. The one pass approach is more straightforward contrasted with the multi-pass approach on the grounds that (a) there is no compelling reason to explicitly recognize the language and (b) utilize a few language-explicit ASRs. In any case, one pass way to deal with ML-ASR presents issues as a need to gather adequate measure of blended language discourse corpus (sound and the related content translation) which can be utilized to construct the blended language acoustic and the ML language display

required for ML-ASR (Mixed Language- Automatic Speech Recognition). In this paper, we theorize that one could utilize accessible assets (for instance acoustic models of one of the dialects in the blended language) and cautiously build the LM and PL to complete a ML-ASR (Mixed Language- Automatic Speech Recognition). We led a few examinations on blended language discourse where the essential language is Hindi and the optional language is English. It ought to be noticed that the methodology is autonomous of the language blend as in some other Indian language can replace Hindi with suitable mapping of the telephonic discussion within the context of the Indian language to the other multilingual blended language also. The objective of programmed discourse acknowledgment (ASR) framework is to precisely and effectively convert a discourse motion into an instant message free of the gadget, speaker or the earth. When all is said in done, the discourse flag is caught and pre-handled at front-end for highlight extraction and assessed at back-end utilizing the Gaussian blend shrouded Markov demonstrate. In this factual methodology since the assessment of Gaussian probabilities command the absolute computational burden, the proper choice of Gaussian blends is essential relying on the measure of preparing information. As the little databases are accessible to prepare the Indian dialects ASR framework, the higher scope of Gaussian blends (for example 64 or more), typically utilized for European dialects, can't be connected for them. This paper audits the measurable structure and shows an iterative methodology to choose an ideal number of Gaussian blends that displays most extreme exactness with regards to Hindi discourse acknowledgment framework. Subsequently, the below diagram depicts the general framework of Automatic Speech Recognition System for perusal and ready reference.

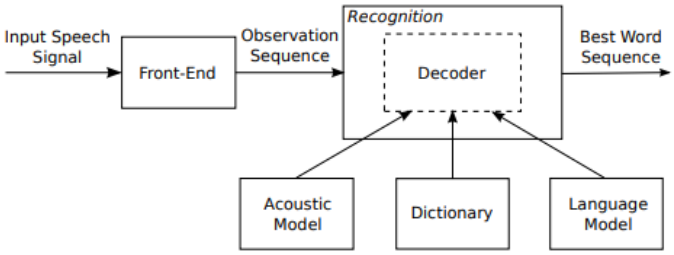

Figure 1 : Automatic Speech Recognition System incorporating Input Speech Signal, Observation Sequence, Deoder based on Acoustic Model, Dictionary and Language Model.

Feature extraction: The information speech signal flag S is normally time-area tested speech waveform. Be that as it may, human hearing depends on the attributes of speech sounds in the recurrence space, in this way a phantom portrayal of speech flag is progressively helpful for speech acknowledgment. Since speech flag is a period differing signal, which is stationary inside a brief timeframe yet changes over a more extended time [Rabiner and Juang, 1993]. While extricating highlights, we have to portion the information speech motion into little edges, at that point procedure each casing independently. The edge length is typically 25 msec. It is short enough to catch the fast changes in speech and adequate to accomplish adequate time-area goals. As the mel-scale approximates the human sound-related reaction better, the Gaussian Mixture Model (GMC) is a standout amongst the most well known component portrayals in speech acknowledgment [Davis and Mermelstein, 1980].

Recognition: Following feature extraction, the recognition component decodes the most probable word sequence W from the observation sequence O. This recognition process can be represented by the following equation:

$$\hat{W} = \arg\max_{W} P(W|O) = \arg\max_{W} \frac{P(W)P(O|W)}{P(O)},$$

where P(W) is the prior probability of the word sequence W, P(O|W) is the likelihood of the observation sequence O given the word sequence W, and P(O) is the probability of observing O. Since P(O) is not a variable of W, Equation can be written as:

$$\hat{W} = \arg\max_{W} P(W)P(O|W).$$

Although the true distribution of P(O|W) and P(W), those probabilities can be estimated from the predefined acoustic model and language model.

Acoustic model: Most ASR frameworks embrace the Hidden Markov models (HMMs) [Baum and Petrie, 1966; Baum and Egon, 1967] to catch the acoustic qualities of discourse sounds. Figure 2 demonstrates the run of the typical topology of HMMs utilized in discourse acknowledgment. The model has three concealed states linked from left to right. In the wake of going into an express, an example can either stay in that state for some time or travel to the following state. The perception of succession O is created by each state and just relies upon that state. To train a HMM, we need to estimate the initial state distribution $\pi = \{\pi_i = P(q_1 = Si)\}$, the transition probability matrix $A = \{a_{ij} = P(q_{t+1} = S_j \,|qt = S_i)\}$ and the observation distribution $B = \{b_i(O_t) = P(O_t \,|q_t = S_i)\}$, where $O = \{O_1, O_2, \cdots, O_T\}$ is a T-frame

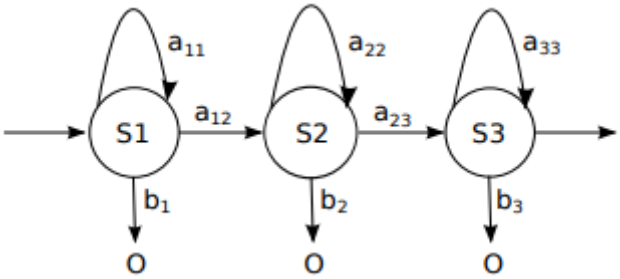

Figure 2: Acoustic Model Framework for Discourse Sounds using Hidden Markov Model vide Speech State Distribution.

observation sequence and $Q = \{q_1, q_2, \cdots, q_T\}$ is the underlying state sequence. The Gaussian mixture model (GMM) is usually used to approximate the observation distribution B, hence the likelihood P(O|W) of the observation sequence O given W can be calculated as:

$$P(O|W) = \sum_{allQ} P(O|Q,\lambda)P(Q|\lambda)$$
$$= \sum_{q_1,q_2,\cdots,q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T).$$

However, HMM parameters λ = (π, A, B) can be estimated using the well known Baum-Welch (BW) algorithm [Baum et al., 1970], a special case of the classical Expectation-Maximization (EM) algorithm [Dempster et al., 1977].

HMM can be prepared on various units, for example, telephones, syllables, words, and so on. As there are less one of a kind telephones than words in a language, preparing telephone HMMs requires considerably less preparing information than preparing word HMMs. Then again, on the grounds that co-verbalizations frequently show up in consistent discourse, the discourse flag of a telephone can be vigorously impacted by encompassing telephones. Just preparing a HMM for each telephone isn't adequate to show the acoustic properties of discourse sounds in various settings. Thus, in discourse acknowledgment, HMMs are typically prepared on triphone, which is a telephone unit gone before and pursued by explicit telephones. In any case, indeed, even simply preparing triphone HMMs, there are still such a large number of triphone units to work with. For model, in our English ASR framework, there are just 39 telephones, however up to 393 = 59319 special triphones. Accordingly, to diminish the measure of preparing information, we bunch triphone HMM states or on the other hand Gaussian blends into gatherings and utilize the information from each gathering for preparing [Hwang, 1993; Huang, 1989].

Language model: The language model is utilized to ascertain the earlier likelihood P(W) of watching the word grouping W in a language. In speech acknowledgment, the language demonstrates is exceptionally useful to separate acoustic questionable speech sounds and lessen the pursuit space amid interpreting. For instance, it is hard to segregate the accompanying two expressions, "SANTA BANTA" and "CENTA BENTA", utilizing acoustic properties. In any case, from our earlier information of English, we realize that the main articulation is bound to hear that the second expression, all things considered. Scientifically, P(W) can be decayed as

$$P(W) = P(w_1, w_2, \cdots, w_n)$$
$$= P(w_1)P(w_2|w_1)\cdots P(w_n|w_1, w_2, \cdots, w_{n-1})$$
$$= \prod_{i=1}^{n} P(w_i|w_1, w_2, \cdots, w_{i-1}),$$

which is a product of the probabilities of observing word w_1 given is defined under the ordinary speech models.

Dictionary: We had talked about the acoustic model and language model in previous sectionss. As appeared in Figure 1, there is another module in the decoder, the word reference. Acoustic model estimates the acoustic properties of speech sounds. Language show assesses the earlier likelihood of word groupings in a language. While the word reference conquers any hindrance between acoustic model and language show with the lexical information. Lexicon gives elocutions of words, so decoder knows which HMMs to use for a specific word. Lexicon additionally gives a rundown of words to restrict the language show intricacy and the decoder's inquiry space. Thus, an ASR framework can just perceive a predetermined number of words displayed in the lexicon, which is regularly known as shut vocabulary search acknowledgment. Table 1. shows some portion of the lexicon utilized in ASR framework. We can find that for certain words, for example, "A", numerous elocutions are given in the lexicon, as there are typically a couple of various approaches to articulate those words.

An example of the dictionary used in our ASR system.

A	AH
A(2)	EY
ABANDON	AH B AE N D AH N
⋮	⋮
INK	IH NG K
⋮	⋮
ZURICH	Z UH R IH K

Table 1. Lexicon Utilization under ASR Framework

It is difficult to create a lexicon without any preparation. To acquire a lexicon explicitly for speech acknowledgment, it generally includes with different etymologists physically compose principles and check singular articulations. This procedure can be in all respects exorbitant and tedious. Not specify that numerous language specialists may not concur with one another and an etymologist may not be steady over a significant lot of time. Analysts had explored to anticipate elocutions of new words with models prepared from existing lexicons [Chen, 2003; Bisani and Ney, 2008]. There are additionally some works on refining a current word reference with expressed models [Bahl et al., 1991; Maison, 2003]. Be that as it may, most word references utilized in ASR frameworks still require human mediation. The extent of a lexicon, i.e., the quantity of one of kind words it contains, is an imperative parameter for an ASR framework. For some area explicit applications, a 5k-word lexicon might be sufficient. For a huge vocabulary persistent speech acknowledgment framework, a 64k-word or bigger lexicons are typically connected. While for voice look frameworks, it is exceptionally basic to apply a lexicon with more than 100k words. An exceptionally substantial word reference may make a few issues an ASR framework. To start with, it requires more information for preparing the acoustic model and language demonstrate, which will create bigger models with more parameters. Accordingly, the decoder will devour more memory to stack those models amid unravelling. Second, a bigger word reference will in general increment the perplexity of the language model to the testing information, which will influence the speed and exactness of the recognizer, since it expands the measure of the pursuit space amid deciphering. Along these lines, we can't generally utilize a vast lexicon for all speech acknowledgment applications

II. LITERATURE REVIEW

S. Itahashi, S. Makino, K. Kido [1] depicts that, the discrete-word recognition system utilizing a word dictionary and phonological rules is described. In this system, nine distinctive features are extracted from a discrete-word input. Segmentation is performed using these features. Segmentation errors are corrected by applying a phoneme connecting rule. The input word is transformed into an input feature matrix. The comparison of this matrix with the standard derived from the dictionary is performed in the feature (matrix) space. Another method of segmentation is also described in which segmentation is performed using a duration dictionary. The effectiveness of utilizing a word dictionary and phonological rules in automatic discrete-word recognition is discussed.

Tilo Sloboda, Alex Waibel [2] depicts that, Spontaneous speech adds a variety of phenomena to a speech recognition task: false starts, human and nonhuman noises, new words, and alternative pronunciations. All of these phenomena have to be tackled when adapting a speech recognition system for spontaneous speech. In this paper we will focus on how to automatically expand and adapt phonetic dictionaries for spontaneous speech recognition. Especially for spontaneous speech it is important to choose the pronunciations of a word according to the frequency in which they appear in the database rather than the \correct" pronunciation as might be found in a lexicon. Therefore, we proposed a data-driven approach to add new pronunciations to a given phonetic dictionary in a way that they model the given occurrences of words in the database. We will show how this algorithm can be extended to produce alternative pronunciations for word tuples and frequently misrecognized words. We will also discuss how further knowledge can be incorporated into the phoneme recognizer in a way that it learns to generalize from pronunciations which were found previously. The experiments have been performed on the German Spontaneous Scheduling Task (GSST), using the speech recognition engine of JANUS 2, the spontaneous speech-to-speech translation system of the Interactive Systems Laboratories at Carnegie Mellon and Karlsruhe University

John Eric Fosler-Lussier [3] depicts that, as of this composition, the programmed acknowledgment of unconstrained discourse by PC is laden with mistakes; numerous frameworks decipher one out of each three to have words mistakenly, while people can interpret unconstrained discourse with one blunder in twenty words or better. This high mistake rate is expected to some degree to the poor displaying of elocutions inside unconstrained discourse. This thesis analyzes how elocutions fluctuate in this talking style, and how talking rate and word consistency can be utilized to foresee when more prominent articulation variety can be normal. It incorporates an examination of the connection between talking rate, word consistency, articulations, and blunders made by discourse acknowledgment frameworks. The after effects of these examinations propose that for unconstrained discourse, it might be proper to assemble models for syllables and words that can powerfully change the elocutions utilized in the discourse recognizer dependent on the all-encompassing setting (counting encompassing words, telephones, expressing rate, and so forth.). Execution of new articulation models consequently got from the information inside the ICSI discourse acknowledgment framework has demonstrated a 4-5% relative enhancement for the Broadcast News acknowledgment task. About 66% of these increases can be ascribed to static base form enhancements; adding the capacity to progressively change articulations inside the recognizer gives the other third of the improvement. The Broadcast News task likewise takes into account examination of execution on different styles of discourse: the new elocution models don't help for pre-arranged discourse, however, they give a significant increase to unconstrained discourse. Not exclusively do the consequently learned articulation models catch a portion of the semantic variety because of the talking style, yet they likewise speak to variety in the acoustic model because of channel effects. The biggest improvement was found in the phone discourse condition, in which 12% of the mistakes delivered by the standard framework were amended.

III. PROPOSED SCHEME

In day by day correspondences, a large portion of time, we can effectively recognize Multilingual Blended Speech Recognition words in human discourse. When we hear a Multilingual Blended Speech Recognition word, other than the vulnerability about individual Multilingual Blended Speech Recognition words, we regularly depend on wide relevant proof, particularly syntactic and semantic proof to check whether what we hear is sensible in a specific setting. In this way, we might most likely improve the Multilingual Blended Speech Recognition word location execution by using abnormal state syntactic and semantic confirmations utilizing Gaussian Mixture Model for Non-Dictionary Words. Although we will apply certain logical highlights when constructing the Multilingual Blended Speech Recognition word classifier in the Multilingual Blended Speech Recognition word discovery part utilizing modified lexicon, longer range conditions between words crosswise over sentences and even sections are as yet not investigated. Other than the Multilingual Blended Speech Recognition word classifier, we can likewise endeavor to discover different methods for applying abnormal state syntactic and semantic confirmations. For instance, we can re-score the grids by utilizing a figured language demonstrate utilizing Hindi and English mix which will be worked from different highlights to locate a superior acknowledgment theory after the primary pass translating utilizing Multilingual Blended Speech Recognition. Subsequently, We may then have a superior acknowledgment exactness and identify progressively Multilingual Blended Speech Recognition words. Then again, we probably won't most likely acquire various cases of a Multilingual Blended Speech Recognition word, on the off chance that we update the acknowledgment vocabulary time after time. For this situation, despite the fact that we can distinguish those Multilingual Blended Speech Recognition words, which we will most likely be unable to effectively recoup their composed structure and language show scores. This issue is increasingly essential for a business

framework, where certain occasions may rise and all of a sudden create extreme interest Multilingual Blended Speech Recognition words. In this way, the beneath figure portrays the proposed framework for prepared reference.

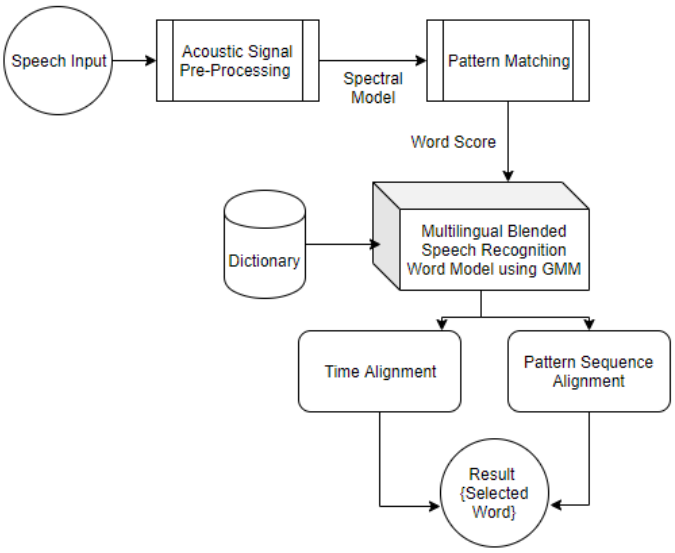

Figure 3: Work Flow of Proposed Scheme comprising of Speech Input, Acoustic Signal Pre-processing, Pattern Matching, Dictionary, GMM Multilingual Model and Results.

REFERENCES

[1] S. Itahashi, S. Makino, K. Kido Discrete-word recognition utilizing a word dictionary and phonological rules IEEE Transactions on Audio and Electroacoustics (Volume: 21 , Issue: 3 , Jun 1973).

[2] Tilo Sloboda, Alex Waibel, DICTIONARY LEARNING FOR SPONTANEOUS SPEECH RECOGNITION, Interactive Systems Laboratories, University of Karlsruhe | Karlsruhe, Germany Carnegie Mellon University | Pittsburgh, USA.

[3] John Eric Fosler-Lussier Dynamic Pronunciation Models for Automatic Speech Recognition,International Computer Science Institute, TR-99-015 September 1999

[4] G. Aradilla, J. Vepa, and H. Bourlard. Using posterior-based features in template matching for speech recognition. In Proc. Interspeech-2006, 2006.

[5] K. Audhkhasi and A. Verma. Keyword search using modified minimum edit distance measure. In Proc. ICASSP-2007, volume 4, pages 929–932, 2007.

[6] L. R. Bahl, S. Das, P. V. deSouza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M. A. Picheny, and J. Powell. Automatic phonetic baseform determination. In Proc. ICASSP-1991,

[7] volume 1, pages 173–176, 1991. L. E. Baum and J. A. Egon. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. Bull. Amer. Meteorol. Soc.,

[8] 73:360–363, 1967. L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics, 37:1554 – 1563, 1966.L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat., 41(1):164–171, 1970.

[9] I. Bazzi and J. Glass. Modeling out-of-vocabulary words for robust speech recognition. In Proc. ICSLP-2000, volume 1, pages 401–404, 2000.

[10] M. Bisani and H. Ney. Open vocabulary speech recognition with flat hybrid models. In Proc. Interspeech-2005, pages 725–728, 2005.

[11] M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication, 50(5):434–451, 2008.

[12] A. W. Black, P. Taylor, and R. Caley. The Festival Speech Synthesis System. University of Edinburgh, 1997.

[13] L. Burget, P. Schwarz, P. Matejka, H. Hermansky, and J. Cernocky. Combining of strongly and weakly constrained recognizers for reliable detection of OOVs. In Proc. ICASSP-2008, pages 4081–4084, 2008.

[14] S. F. Chen. Conditional and joint models of grapheme-to-phoneme conversion. In Proc. Eurospeech-2003, pages 2033–2036, 2003.

[15] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling.Technical report tr-10-98, Center for Research in Computing Technology (Harvard University), August 1998.

[16] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speechand Signal Processing, 28(4):357 – 366, 1980.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc., 39(1):1–38, 1977. C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, 1998.

[18] J. Fiscus, J. Garofolo, M. Przybocky, W. Fisher, and D. Pallett. 1997 English Broadcast News Speech (HUB4). Linguistic Data Consortium, 1998.

[19] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In Proc. ASRU-1997, pages 347–354, 1997.

[20] M. J. F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. Computer Speech and Language, 12:75–98, 1998.

[21] L. Galescu. Recognition of out-of-vocabulary words with sub-lexical language models. In Proc Eurospeech-2003, pages 249–252, 2003.

[22] J. Garofalo, D. Graff, D. Paul, and D. Pallett. CSR-I (WSJ0) Complete. Linguistic Data Consortium, 1993.

[23] J. Garofalo, D. Graff, D. Paul, and D. Pallett. CSR-II (WSJ1) Complete. Linguistic Data Consortium, 1994.

[24] J. J. Godfrey and E. Holliman. Switchboard-1 Release 2. Linguistic Data Consortium, 1997. I. Good. The population frequencies of species and the estimation of population parameters.Biometrika, 40:237–264, 1953.

[25] D. Graff, R. Rosenfeld, and D. Paul. CSR-III Text. Linguistic Data Consortium, 1995.

[26] D. Graff, J. Garofolo, J. Fiscus, W. Fisher, and D. Pallett. 1996 English Broadcast News Speech (HUB4). Linguistic Data Consortium, 1997.

[27] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. volume 1, pages 13–16, 1992.