ACOUSTIC-PHONETIC INFORMATION FROM EXCITATION SOURCE FOR REFINING MANNER HYPOTHESES OF A PHONE RECOGNIZER

Dhananjaya N.

Dept. of Computer Science & Engineering
IIT Madras, Chennai, India
dhanu@cse.iitm.ac.in

B. Yegnanarayana & Suryakanth V. G.

International Institute of Information Technology
Hyderabad, India
{*yegna,svg*}*@iiit.ac.in*

ABSTRACT

Reliable acoustic-phonetic (AP) information derived from the speech signal can be used to detect and correct errors in the output of a phone recognizer. In this paper, limited acoustic-phonetic information derived primarily by processing the excitation source information in the speech signal is used to improve the performance of detection of manner of articulation from a baseline phone recognition system. A context-independent HMM-based monophone system without any language information is used as the baseline system for this purpose. The performance of the phone recognizer in terms of its ability to detect the manners of articulation is studied. The errors in the hypothesis of the manner of articulation of phones are corrected using AP information such as voicing, voice bar and frication. It is shown that significant improvement can be achieved by using simple or limited AP information.

Index Terms— Acoustic-phonetic, excitation source, zero-frequency, manner of articulation, voicing, voice bar, frication

1. INTRODUCTION

Popular systems for phone recognition employ a statistical approach for building phone models, and use the standard spectral features such as mel-frequency cepstral coefficients (MFCCs) [1, 2]. The spectral features do not contain all the necessary information to distinguish between various sounds in speech, and additional information, such as from the excitation source, needs to be used to improve the performance of existing systems [3]. Also, the acoustic-phonetic (AP) features of the phonemes are used to a minimal extent in these systems, with most of the information embedded implicitly in the features used. Explicit and reliable extraction of the acoustic-phonetic features can help in two ways: (a) It can provide an alternative approach for phone recognition [4], and (b) it can be used to improve the performance of existing phone recognizers [2]. Towards this goal it is important to derive the acoustic correlates from the speech signal to obtain the AP attributes of a sound such as the manner and place of articulation. Manner of articulation (MoA) is an important piece of acoustic-phonetic information which is mainly associated with the excitation source of the speech production apparatus. In this paper, we explore the possibility of using reliable AP features derived primarily from the excitation source in correcting errors in a baseline phone recognizer. A context-independent HMM-based monophone system without any language information is used as the baseline system. In particular, we study the performance of the phone recognizer in detecting the manners of articulation of the phones. Acoustic correlates for identifying phonetic attributes namely, voicing, voice bar (voiced closure) and frication are derived from the signal and are used to correct the manner information of the hypothesized phones.

The paper is organized as follows: Section 2.1 describes the baseline phoneme recognition system using monophone HMM models. The performance of the phoneme recognizer in identifying the manner of articulation of the various phones is analyzed. In Section 3, the signal-based evidences are used to detect the AP features such as voiced/nonvoiced, voice bar and frication. These AP features are used for correcting some of the errors in the phoneme recognizer output. Section 4 gives a summary of the paper.

2. MANNER OF ARTICULATION IN CONTINUOUS SPEECH

Manner of articulation is a phonetic attribute of sound which primarily describes the nature of excitation source used during the production of the sound. The list of ten manner labels used, and the mapping of the phone labels to each of the ten manners is given in Table 1. Traditionally, only six manners namely {*silence, stop, fricative, nasal, approximant and vowel*} are used in developing phone recognizers. The manners for voiced and unvoiced closures ([ucl] and [vcl]), along with voiced and unvoiced bursts ([vbt] and [ubt]), are merged into a single manner, denoting a stop. But in continuous speech it is observed that a closure need not always be followed by a burst, and also a significant number of voiced bursts are not clearly articulated or are so weak that they are not manifested clearly in the acoustic signal. It is important

Table 1. List of manners of articulation, and mapping of phone labels to their corresponding manners of articulation. The number of reference phones for each MoA is also given.

Manner		Symbol	Phones	# Ref
Silence		[sil]	[#h], [h#], [sil], [pau], [epi]	530
Stop	Unvoiced closure	[ucl]	[pcl], [tcl], [kcl], [qcl]	826
	Voiced closure	[vcl]	[bcl], [dcl], [gcl], [dx]	603
	Unvoiced burst	[ubt]	[p], [t], [k]	726
	Voiced burst	[vbt]	[b], [d], [g]	412
Fricative	Unvoiced fricative	[ufr]	[s], [sh], [f], [th], [ch], [jh], [hh]	817
	Voiced fricative	[vfr]	[z], [v], [dh], [hv]	491
Nasal		[nas]	[m], [n], [ng], [em], [en]	736
Approximant/ Semivowel		[svl]	[y], [r], [l], [w], [el]	1004
Vowel		[vow]	[iy], [ih], [eh], [ey], [ae], [aa], [aw], [ay], [ah], [ao], [oy], [ow], [uh], [uw], [ux], [er], [ax], [ix], [axr], [ax-h]	2953

to discriminate between voiced and unvoiced closures for discriminating between similar words such as {*band, gant, pant, can't, canned*}. One reason for merging the voiced and unvoiced closures is the inability of most algorithms to detect the weak voicing present during voiced closures (also referred to as a voice bar). Also, phonetically, it is more appropriate to label what is articulated than what is expected. In this paper, we consider all the four manner labels separately. Similarly the unvoiced and voiced fricatives are considered as separate manners.

2.1. Baseline phone recognition system

A context-independent monophone hidden Markov model (HMM) based phone recognizer is used as the baseline system [1]. The system does not use language information in any form. A 3-state left-to-right HMM model with a 64 mixture continuous-density diagonal-covariance Gaussian mixture model (GMMs) per state is used to model each of the phones. The baseline phone recognition system is built for the TIMIT dataset, with the 48 phone labels listed in [1], with the exception that the voiced closures ([bcl], [dcl] and [gcl]) are not merged into a single voiced closure label. The voiced fricative [zh] is mapped or merged in to its unvoiced counterpart [sh], and the label for epithetic silence [epi] is mapped onto the silence label [sil], thereby resulting in a total of 48 phones. The entire TIMIT train dataset (462 speakers, each with 10 short (3-5 s) sentences), is used for building the phone models. Ten iterations of the Viterbi training is followed by an embedded Baum-Welch training for another ten iterations. The open source HMM tool kit (HTK) [5] is used for building the phone recognition system. The performance of the baseline phone recognition system when tested on the TIMIT core test set (24 speakers, each with 10 sentences) is given in Table 2 using the optimal string matching algorithm based on dynamic programing [5]. The performance is given for the standard 48 phones and for the reduced 39 phone set given in [1]. The string matching strategy for performance evaluation depends only on the sequence of phone labels hypothesized, and do not give any insight on the accuracies of the phone boundaries detected. The performance of the system using a word-spotting (rather phone-spotting) strategy, where a reference phone label is considered correct (a hit) if at least 50% of its duration matches that of a hypothesized phone with the same label [5], is also given in the third row in Table 2. Similarly, a hypothesized phone label is considered a false alarm when less than 50% of its duration matches with that of a matching reference label [6].

Table 2. Performance of the baseline phone recognition system using optimal string matching. P_c denotes the percentage of correct detections. *Acc* denotes the accuracy defined as $Acc = P_c - P_i$, where P_i is the percentage of insertion errors. P_h denotes the percentage of correct detections (hits) and P_{fa} denotes the percentage of false alarms.

	String matching		Phone-spotting	
# symbols	P_c (%)	*Acc* (%)	P_h (%)	P_{fa} (%)
48 phones	69.3	56.28	65.5	31.8
39 phones	75.17	61.70	70.1	27.2
10 MoA	81.1	77.7	79.3	17.0

2.2. Manner detection

Manner information can be derived from the hypothesized phone labels by mapping each of the phone label to a corresponding manner of articulation. The performance of the baseline system in detecting the manner of articulation (MoA) of a phone is given in Table 2. The confusion matrix of the baseline phone recognizer in detecting the manner of articulation of the phones is given in Table 3. It can be seen that one of the primary sources of error is in detecting the voicing information correctly, such as between unvoiced and voiced closures, unvoiced and voiced bursts, and unvoiced and voiced fricatives. The objective of this study is to improve the performance of the baseline phone recognizer in hypothesizing the manners of articulation using reliable features from the

Table 3. Performance of the baseline phone recognition system in terms of confusion in detecting the manner of articulation. [del] denotes deletions.

	[sil]	[ucl]	[vcl]	[ubt]	[vbt]	[ufr]	[vfr]	[nas]	[svl]	[vow]	[del]
[sil]	**93.20**	1.13	0.75	0.18	0	0.75	0.56	0.94	0.18	0.56	1.70
[ucl]	3.14	**62.71**	14.16	2.66	0.36	9.07	2.66	1.81	1.69	0.48	1.21
[vcl]	0.66	8.62	**72.13**	0.33	1.16	1.82	3.48	3.98	1.32	4.97	1.49
[ubt]	2.20	2.20	0.68	**53.71**	7.16	11.01	2.47	1.51	2.06	16.39	0.55
[vbt]	0.97	0.24	3.39	10.19	**67.71**	2.66	4.85	1.45	5.09	3.39	0.00
[ufr]	0.61	1.34	0	4.03	0.48	**87.02**	5.26	0.12	0.12	0.12	0.86
[vfr]	0.61	0.40	2.64	2.03	3.46	21.99	**60.48**	3.25	1.01	3.46	0.61
[nas]	0	1.76	5.97	0	0.27	0.67	0.81	**80.70**	3.12	6.11	0.54
[svl]	0	0	0.19	0.09	0.29	0.69	1.49	1.19	**72.21**	23.60	0.20
[vow]	0.06	0	0.40	0.13	0.03	0.33	0.33	1.32	5.68	**90.51**	1.15

excitation source.

3. ACOUSTIC-PHONETIC INFORMATION USING EXCITATION SOURCE FEATURES

The acoustic-phonetic features explored in this paper, namely, voicing, voice bar and frication, rely predominantly on excitation source features derived from the speech signal. Excitation source features, namely, the instants of glottal closure (epochs), strength of excitation at the epochs and instantaneous fundamental frequency or pitch are derived from the zero-frequency analysis of the speech signal [7]. Zero-frequency analysis of speech is motivated by the fact that excitation source is impulse-like (for voiced speech), and its effect is felt throughout the spectrum, including at zero frequency, where the effect of vocal tract is minimal. Normalized error for different linear prediction orders, normalized zero-frequency filtered (ZFF) signal strength are some of the other excitation source features used [8]. Vocal tract system features such as dominant resonance frequency and it strength derived from the modified group delay spectrum is used along with the excitation source features [9].

3.1. Voicing features

The strength of excitation as measured from the ZFF signal is a good measure of voicing, even in weak voiced phones such as voiced closures [10]. In [11] a voiced/nonvoiced (V/NV) detection algorithm is proposed, which utilizes the robustness of the epoch locations as well as the strength of excitation measured from the ZFF signal. The performance of the phone recognition system in terms of the confusion in identifying the manner of voicing of phones is given in Table 4. The improvement in performance after validating the voicing decisions hypothesized by the phone recognizer using the V/NV evidence obtained using the excitation source is also given. A phone hypothesized as voiced by the phone recognizer is considered valid, if more than 50% of the phone duration is identified as voiced using the excitation source information. A similar validation is also performed on the nonvoiced phones. It is seen that the overall voiced/nonvoiced detection accuracy improves from about 90% to around 95%.

Table 4. Performance of the baseline phone recognizer in detecting the manner of voicing of phones, before and after correction using AP features from excitation source.

	Before		After	
	[NV]	[V]	[NV]	[V]
[NV]	90.4	9.6	95.3	4.7
[V]	10.9	89.1	5.8	94.2

3.2. Voice bar

Voice bars are regions of closure during the production of voiced stop consonants. These are regions of weak voicing and most voicing detection algorithms fail to detect or ignore. The performance of the phone recognizer in detecting the regions of voiced closure is given in Table 5. The symbols [OUVP] and [OVP] denote other unvoiced and voiced phones respectively. The voice bars have the highest confusion with unvoiced closures, and a large number of confusions with other voiced phones is due to nasals and the voiced fricative [dh], which have features closest to voice bars, especially in continuous speech. In [8] we have proposed a knowledge-based system for detection of the voice bars in continuous speech. Excitation source features namely excitation strength, normalized ZFF signal strength, normalized linear prediction error, along with dominant resonance frequency (vocal tract system feature derived from the phase spectrum) are used to detect the voice bars. Every phone labeled by the baseline system as a voiced closure ([vcl]) is validated, if at least 50% of the segment is detected as voice bar using the AP features. Similarly, phones hypothesized as [ucl], [nas] and [vfr] are validated using the voice bar evidence derived from the signal. The improvement in performance is given in Table 5. It is to be noted that the validations performed improves the overall recognition accuracies of nasals and voiced fricatives.

3.3. Frication

The performance of the baseline system in spotting phones with frication as a manner of excitation is given in Table 6. The magnitude of the modified group delay (MGD) function [9] at zero frequency is used to validate a phone hypothesized

Table 5. Performance of the baseline phone recognizer in spotting phones with voiced closure before and after using the AP evidence.

	Before				After			
	[OUVP]	[ucl]	[vcl]	[OVP]	[OUVP]	[ucl]	[vcl]	[OVP]
[OUVP]	83.6	1.6	0.4	13.3	83.6	1.6	0.4	13.3
[ucl]	14.9	62.7	14.2	7.0	14.9	70.5	8.5	4.8
[vcl]	2.8	8.6	72.1	14.9	2.8	4.8	86.2	8.0
[OVP]	3.7	0.3	1.5	93.7	3.7	0.3	0.9	94.4

Table 6. Performance of the baseline phone recognizer in spotting phones with frication before and after using the AP evidence.

	Before				After			
	OUVP	[ufr]	[vfr]	OVP	OUVP	[ufr]	[vfr]	OVP
OUVP	71.5	7.6	2.1	17.7	75.6	4.5	1.1	17.7
[ufr]	6.0	87.0	5.3	0.9	6.0	89.7	2.6	0.9
[vfr]	3.1	22.0	60.5	13.9	3.1	11.0	71.5	13.8
OVP	2.2	0.8	1.3	94.9	2.2	0.4	1.3	95.3

as unvoiced fricative. The magnitude of the MGD is positive for voiced sounds, while it is negative for regions of frication. A good number of voiced fricatives (especially [z] and [zh]) are normally hypothesized as unvoiced fricatives. This can be reduced by using the voicing evidence derived from the ZFF signal and the frication evidence derived from the MGD. A phone labeled as an unvoiced fricative is corrected as voiced fricative, if there is an overlap of voicing and frication evidence for at least 10% of the phone duration. The overall improvement in performance of detection of frication is given in Table 6.

4. CONCLUSIONS

In this paper, the acoustic-phonetic information derived primarily from the excitation source information in the speech signal was used to correct errors in the manner hypotheses of a baseline phone recognizer. It was shown that significant improvement can be achieved by using simple or limited AP information derived from the excitation source of the speech signal. The overall performance of manner detection has improved from 79.3% to 86.2%. The overall phone recognition performance of the baseline system can be improved, if a similar correction of errors can be made on the hypotheses of place of articulation. We are currently working on correcting errors in the hypotheses of place of articulation using AP features that can be derived from the modified group delay spectrum. An improved hypotheses of acoustic-phonetic attributes can be used to develop an alternate phone recognizer along the lines outlined in [4].

5. REFERENCES

[1] Kai-Fu Lee and Hsiao-Wuen Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoustics Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1989, Nov. 1989.

[2] S. M. Siniscalchi and Chin-Hui Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, pp. 1139–1153, 2009.

[3] Bishnu S. Atal, "Automatic speech recognition: A communication perspective," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, Phoenix, AZ, USA, March 1999, pp. 457–461.

[4] Sharlene A. Liu, "Landmark detection for distinctive feature-based speech recognition," *Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, Nov. 1996.

[5] Steve Young, et. al., *The HTK Book (for HTK version 3.4)*, Cambridge University Engineering Department, Cambridge, UK, 2009.

[6] Jinyu Li and Chin-Hui Lee, "On designing and evaluating speech event detectors," in *Proc. Interspeech*, Lisbon, Portugal, Sept. 2005, pp. 3365–3368.

[7] K. Sri Rama Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[8] Dhananjaya N, S. Rajendran, and B. Yegnanarayana, "Features for automatic detection of voice bars in continuous speech," in *Proc. Interspeech*, Brisbane, Australia, Sept. 22-26 2008, pp. 1321–1324.

[9] Anand Joseph M, Guruprasad S, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. Int. Conf. Spoken Language Processing (INTERSPEECH)*, Pittsburgh PA, USA, Sept. 2006, pp. 1009–1012.

[10] K. Sri Rama Murty, B. Yegnanarayana, and M. Anand Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, June 2009.

[11] Dhananjaya N and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, March 2010.