CALIBRATION AND FUSION OF QUERY-BY-EXAMPLE SYSTEMS - BUT SWS 2013

Igor Szöke, Lukáš Burget, František Grézl, Jan "Honza" Černocký and Lucas Ondel*

BUT Speech@FIT, Brno University of Technology, Czech Republic
{szoke,burget,grezl,cernocky,iondel}@fit.vutbr.cz

ABSTRACT

This paper summarizes our work for MediaEval 2013 Spoken Web Search task evaluations. The task was Query-by-Example (search of spoken queries within spoken data). We submitted a system composed of 26 subsystems, of which 13 are based on Acoustic Keyword Spotting and 13 on Dynamic Time Warping. All of them use three-state phoneme posteriors as input features. Our main contribution was *m-norm* normalization of particular subsystems together with the fusion based on binary logistic regression. The results, including per-language analysis, are provided on MediaEval 2013 dataset.

Index Terms— query-by-example spoken term detection, acoustic keyword spotting, dynamic time warping, fusion, z-norm, m-norm, TWV

1. MOTIVATION AND SYSTEM OVERVIEW

Spoken Web Search task at MediaEval 2013 (*SWS2013*) [1] was the third in a series aiming at Query-by-Example Spoken Term Detection (*QbE*). The goal of Spoken Web Search task (SWS) is to search for audio query within audio content.

As shown in a recent summary paper [2], the QbE approaches can be roughly divided into two categories: the pattern-matching approaches look for similarities at the feature level and are mostly represented by Dynamic Time Warping (DTW)-style comparison of query and utterance segments. The second category is represented by Acoustic Keyword Spotting (AKWS) that builds a model of query and processes the utterances by looking at log-likelihood ratio of the keyword model and a background model.

In our previous research aiming at SWS2012 evaluations [3], we found AKWS superior to DTW in a situation where we were able to build QbE system in a supervised manner (phonetic transcript was provided for development utterances, so that a phone-state posterior estimator could be trained on them). We also found, that DTW based on phoneme posteriors has significant drawback when going from in-language to cross-language condition [4]. According to the above facts, we believed, that DTW is sub-optimal, especially in multilingual and zero-resource conditions in challenging acoustic conditions.

The SWS2013 dataset was challenging in the way of mixed language and acoustic conditions, and we wanted to thoroughly compare and combine DTW and AKWS approaches. In addition, we improved the Artifitial Neural Network (ANN) based phoneme-state estimators by unsupervised ANN adaptation and, inspired by the work of our group in speaker and language identification, we cared about proper normalization and fusion of several systems.

Our **Query-by-Example** (QbE) system (figure 1) is based on phoneme-state posterior estimators. Each estimator (denoted as

* Igor Szöke was supported by Grant Agency of Czech Republic post-doctoral project No. GP202/12/P567.

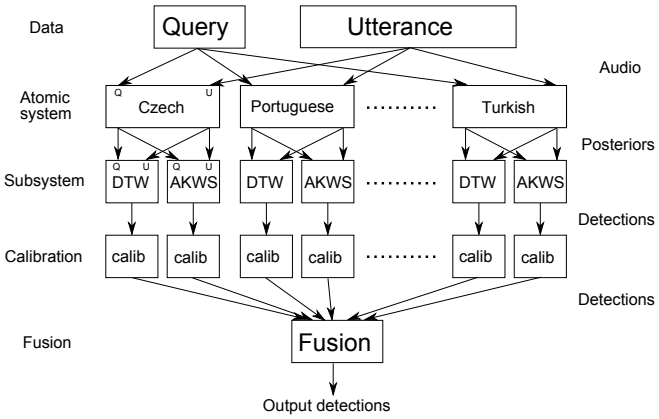

Fig. 1. *Our Query-by-Exmaple system schema. On the right side is the type of the data exchanged among the boxes. On the left side is type of the boxes. Q means Queries as the input, U means Utterances as the input.*

Atomic system) is an artificial neural network taking raw audio file as the input (either query example or test utterance) producing phoneme state posteriors as the output. We used 13 Atomic Systems, see section 3 for details.

Phoneme state posteriors were then processed by **Query-by-Example Subsystems**. We have two types of subsystems, one based on AKWS (section 4) and the other based on DTW (section 5). The input of each subsystem is the matrix of phoneme state posteriors for query example and utterance. The output is a set of detections of given query example in the utterance.

The next step is subsystem **score normalization and calibration**. It takes the set of detections and normalize the detection scores with respect to the normalized cross entropy. Section 6 discusses our findings in details.

Fusion is the final stage of the QbE system. It takes normalized outputs of all subsystems and fuses them into one output. Again, we optimize the fusion parameters with respect to the normalized cross entropy (see section 7).

2. DATA AND SCORING

The SWS2013 development/evaluation database consists of only one set of utterances – used both for development and evaluation – and two sets of queries: one for development and the other for evaluation. The overall length of utterance data is 20 hours. There are 9 languages in 6 subsets (see table 1).[1]

[1]The reader is kindly referred to visit http://speech.fit.vutbr.cz/software/sws-2013-multilingual-database-query-by-example-keyword-spotting for more

	minutes/segments	dev/eval	type
Isixhosa	65/395	25/25	read
Isizulu	59/395	25/25	read
Sepedi	69/395	25/25	read
Setswana	51/395	25/25	read
Albanian	127/968	50/50	read
Basque	192/1841	100/100	broadcast
Czech	252/3667	94/93	conversational
NNEnglish	141/434	61/60	lecture
Romanian	244/2272	100/100	read
SUM	1196/10762	505/503	mixed

Table 1. *The upper part is set of 4 South African languages, the bottom part is set of 5 European languages. The first column: amounts of data per language. The second column: the numbers of development (dev) and evaluation (eval) queries. The last column is type of speech.*

Given that utterances in the search repository were shuffled and no side information was provided to participants regarding the spoken language or the acoustic conditions, any adaptation needs to rely on unsupervised algorithms, thereby introducing an interesting line of research.

Note that the 9 languages selected for this database mostly cover European and African language families. In addition, the non-native English database consists of a mixture of native and non-native English speakers presenting their oral talks. This introduces a **significant pronunciation mismatch** between utterances, as this subset includes utterances with, for example, strong Indian, French, Chinese and other accents. Another interesting aspect of this database is the **variety of speaking styles** (read versus spontaneous (Czech) versus lectures (NNEnglish)) and the variety of matched/**mismatched query-utterances conditions**, which forces us to build system with low/zero resources constraints. A clear example can be found in the Basque subset with queries recorded in isolation by mobile phone in order to retrieve utterances recorded from a broadcast news TV channel.

We report our results in terms of Actual/Maximum Term Weighted Value (ATWV/MTWV) and Upper Bound Term Weighted Value (UBTWV). The UBTWV finds the best threshold for each term separately (it maximizes the TWV for the term) and then calculates the overall TWV [5]. It can be understand as oracle TWV or TWV for a system having ideal score normalization. More details on evaluation metrics used for SWS2013 can be found in [6].

3. ATOMIC SYSTEMS

All our Atomic systems use Artificial Neural Network classifiers (ANN) to estimate per-frame phoneme state posterior probabilities (so-called posteriorgrams). Our motivation was to re-use as many already trained phoneme posterior estimators (Atomic systems) available at Brno University of Technology (BUT) as possible.

The ANNs were trained as acoustic models for phoneme or LVCSR recognizers in several past or running BUT projects[2]. Altogether, we ended-up with 13 Atomic systems with the following architectures and trained on the following datasets:

- $3\times$ **SpeechDat**[3] (Czech, Hungarian and Russian; monolingual LCRC systems [7], trained on 20 hours of read speech per language),
- $1\times$ **BABEL** (Cantonese, Pashto, Tagalog, Turkish; multilingual stacked-bottleneck system [8], 100 hours of conversational speech per language). The BABEL ANN is one network trained on 4 languages using split softmax approach – each language has a separate part of the output layer with its own softmax. The overall number of ANN outputs is 660.
- $1\times$ **SWS2012** (MediaEval SWS2012 development data – isiNdebele, Siswati, Tshivenda, and Xitsonga); multilingual stacked-bottleneck system [3], 1 hour of read speech per language). The SWS2012 ANN is one networked trained on 4 languages having common phoneme set (IPA).
- $8\times$ **GlobalPhone** (Czech, English, German, Portuguese, Russian, Spanish, Turkish, Vietnamese; monolingual stacked-bottleneck systems [9, 10], 20 hours of read speech per language).

We also used unsupervised ANN adaptation on SWS2013 data. We labeled the data with phoneme state labels using decoding with free phoneme loop of the particular language. Then we retrained the GlobalPhone and BABEL ANNs from scratch on the SWS2013 data using the generated state alignments as the ANN targets. We saw overall general improvement of TWV for both DTW and AKWS subsystems (maximum improvement from MTWV 0.1521 to 0.2183 for Portuguese GP DTW subsystem). Detailed analysis is beyond the scope of this paper and is available from author's web pages[4].

4. ACOUSTIC KEYWORD SPOTTING BASED QBE

The Acoustic Keyword Spotting (AKWS) based Query-by-Example subsystems follows our paper [11]. We built an HMM for each query and then calculated log likelihood ratio between the query model and a background model (free phone loop). In QbE task, however, we need to generate the phoneme sequence for each of the acoustic examples – **query-to-text step**. This is achieved by decoding each example using free phoneme loop. We cut-off initial and final silence labels (if present) and omit queries having less than three non-silence phonemes, as these short queries could generate huge amounts of false alarms.

5. DYNAMIC TIME WARPING BASED QBE

In our implementation, we follow the standard Query-by-Example recipe – subsequence DTW [12]. A single DTW is run for each combination of query and utterance and the query is allowed to start at any frame of the utterance. When selecting the locally optimal path in the standard DTW algorithm, transition from the smallest accumulated distance is chosen. In our implementation, we compare the accumulated distances (including the current local distance) normalized by the corresponding path lengths on-the-fly. Note that in the standard subsequence DTW, no on-the-fly path length normalization in performed, which results in the inappropriate preference for shorter (recently started) paths. As the distance metric, we used the usual negative logarithm of the dot product of phone-state posterior vectors.

details about the SWS2013 database and for further references.

[2]Please bear in mind that reusing all these Atomic systems leads to many inconsistencies among them — feature extraction, sizes and structures of ANNs, etc.

[3]http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

[4]http://merlin.fit.vutbr.cz/wiki/index.php/SWS2013QbE

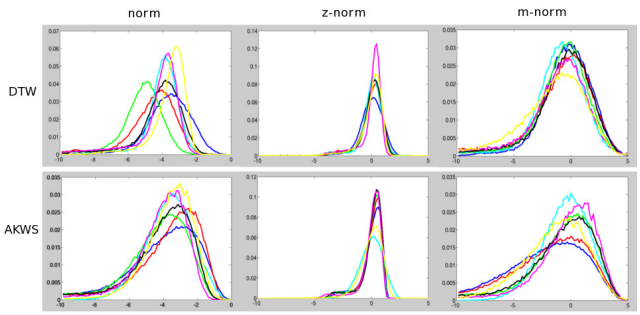

Fig. 2. *Distributions of scores of 6 randomly chosen terms for AKWS and DTW sub-systems based on GlobalPhone Portuguese Atomic system. Norm – raw score of detection is divided by the length of the detection. Z-norm – mean and variance normalization is applied on top of norm. M-norm – m-norm is applied on top of norm.*

	AWKS		DTW	
	MTWV	UBTWV	MTWV	UBTWV
norm	0.0005	0.1987	0.0000	0.3231
norm-sideinfo	0.0826	0.2000	0.0746	**0.3282**
z-norm	0.0007	0.1808	0.0775	0.3094
z-norm-sideinfo	0.0557	0.1817	0.1729	0.3091
m-norm	**0.1162**	0.2078	**0.2138**	0.2886
m-norm-sideinfo	0.1098	**0.2100**	0.1908	0.3053

Table 2. *MTWV (Maximum Term Weighted Value) and UBTWV (Upper-Bound Term Weighted Value) of SWS2013 development queries for GlobalPhone Portuguese atomic system. We compared both AKWS and DTW subsystems for various score normalization techniques. The* sideinfo *tag means that the scores were calibrated using approach described in section 6.1.*

We further improved the DTW systems by applying Voice Activity Detector (VAD) to cut-off the initial and the final silence from the query examples. This improved the overall DTW system by 10% relative.

6. SCORE NORMALIZATION

For both DTW and AKWS subsystems, the local maxima of frame-by-frame accumulated detection scores are selected as candidate detections. For overlapping detections, only the best scoring ones are preserved. For AKWS, the accumulated detection scores are normalized by the length of the detection, for DTW, by the length of warping path (done on-the-fly). These scores are denoted as *norm*.

There might be significant differences between the score distributions corresponding to different queries and it is important to normalize the scores for each query to allow for a single common threshold maximizing the TWV metric (figure 2).

We adopted two normalization approaches: *z-norm* was demonstrated as a promising normalization for QbE in SWS2012 evaluations [13]. It applies mean and variance normalization of scores for each query separately. The variance and mean should be estimated on non-target detections (false-alarms). In case of QbE, the number of non-targets is larger in orders of magnitude than the number of targets (true hits), therefore, we can calculate the mean and variance on the whole set of detection scores (both targets and non-targets) without any significant difference.

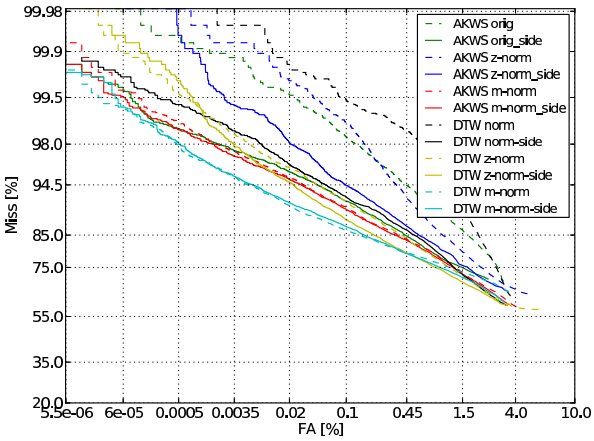

Fig. 3. *DET (Detection Error Trade-off) curves of SWS2013 development queries for GlobalPhone Portuguese atomic system. AKWS and DTW subsystems are compared with various score normalization techniques (TWVs are in table 2). Note cyan and red curves for m-norm. Solid lines represent* sideinfo *calibrations.*

m-norm is our new score normalization technique we proposed for SWS2013 evaluations. It is motivated by the observation that score distributions have very long tails towards small scores, which significantly differ in shape from query to query (figure 2). m-norm is similar to z-norm, but when estimating the variance, it takes into account only subset of the scores. First, maximum of the score distribution (approximated by histogram) is found . Let us denote the score value corresponding to the maximum in distribution of particular query q as $peak_{score}(q)$. Then we estimate standard deviation (denoted $peak_{stddev}(q)$) of set of query scores greater than $peak_{score}(q)$. Finally, we subtract the $peak_{score}(q)$ to shift the peak to 0 and then we divide the scores of all query detections by $peak_{stddev}(q)$.

We evaluated MTWV and UBTWV on development queries for Portuguese GlobalPhone Atomic system and corresponding AWKS and DTW subsystems (table 2), and also plot corresponding DET curves (figure 3). Note that similar behavior was observed also for other subsystems.

6.1. Calibration Using Side Information

Next, we calibrated the scores using binary logistic regression (the same as we used in fusion in section 7), where the input to the logistic regression was a vector of norm, z-norm, or m-norm scores augmented with different per-term side-information scores [14] – denoted as *sideinfo*. The best tested side information, which significantly improved MTWV, was *the logarithm of the number of detections* of a particular term.

According to the results (TWVs - table 2, DET curves - figure 3), the calibration using sideinfo significantly helped in case of *norm* scores for both AKWS and DTW subsystems. Sideinfo helped for the z-norm scores on AKWS subsystem only. With m-norm, the sideinfo improves neither MTWV nor DET curve. This leads to conclusion that z-norm is not sufficient to properly normalize score distributions over different queries and the information about *the number of term detections* can help to make useful correction (shift) to the distribution. After application of m-norm, the scores are already well normalized

Approach	ALL	Albanian	**Basque**	**Czech**	**NNEnglish**	**Romanian**	Isixhosa	Isizulu	Sepedi	Setswana
AKWSDTW	0.3776	0.5969	0.2989	0.1194	0.0601	0.6291	0.4780	0.3983	0.4517	0.2695
DTW	0.3557	0.5889	0.2403	0.1150	0.0594	0.5812	0.5261	0.4192	0.4860	0.3221
AKWS	0.3041	0.4460	0.3195	0.0993	0.0752	0.5299	0.4301	0.2971	0.3879	0.2127

Table 3. *MTWV results of systems on evaluation queries for whole data set (ALL) and for 9 particular languages. DTW denotes fusion of 13 DTW subsystems, AKWS denotes fusion of 13 AKWS subsystems and AKWSDTW denotes fusion of 26 AKWS and DTW subsystems. 4 interesting conditions are typeset in bold: Basque – query and utterance mismatch, Czech – conversational speech, NNEnglish – dialect in query and utterance mismatch, Romanian – read speech, acoustic match of query and utterance.*

Subset	eval			dev		
Approach	ATWV	MTWV	UBTWV	ATWV	MTWV	UBTWV
AKWSDTW	0.3751	0.3776	0.4835	0.4349	0.4373	0.5310
DTW	0.3535	0.3557	0.4585	0.4180	0.4199	0.5153
AKWS	0.3003	0.3041	0.4165	0.3642	0.3644	0.4713

Table 4. *Results for our query-by-example approaches in Actual TWV, Maximum TWV and Upper Bound TWV for development (dev) and evaluation (eval) queries. DTW denotes fusion of 13 DTW subsystems, AKWS denotes fusion of 13 AKWS subsystems and AKWSDTW denotes fusion of 26 AKWS and DTW subsystems.*

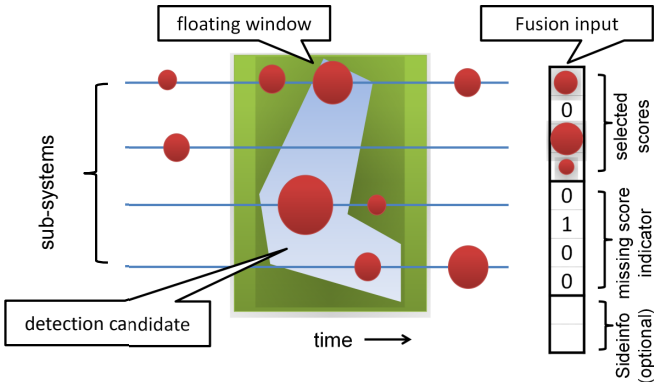

Fig. 4. *System combination of different keyword spotting systems, including alignment and filtering step, as well as creation of feature vector for logistic-regression classifier (from [14]).*

7. FUSION

Normalized scores from the individual subsystems were fused following [14]. The scores from different subsystems are first aligned in time and then linearly combined. The alignment respects a floating window in which all putative hits are expected to represent particular occurrence of reference detection (figure 4). If a subsystem provides more putative hits, only the one with the highest score is considered. In case no putative hit is declared by a subsystem, a zero score is used for the particular system and the corresponding missing score indicator (figure 4) is set to 1. The aligned scores together with the missing score indicators (and optionaly *sideinfo*) form a vector representing one candidate detection. Such vectors are used as the input to binary logistic regression linear classifier, which is trained on the development data and used to produce fused scores for evaluation data.

8. RESULTS

Table 4 summarizes several systems submitted to SWS2013 evaluations. *AKWSDTW* is the fusion of 26 systems while *DTW* or *AKWS* are fusions of 13 DTW or AKWS subsystems respectively. It is worth to note, that DTW achieves significantly better accuracy than AKWS. However both systems are complementary.

In table 3, we analyzed per language accuracy. As can be seen, read speech subsets achieved very good accuracy (except Setswana). On the other hand Czech, NNEnglish and Basque achieved worse results, that, in our opinion, are due to: significant acoustic condition mismatch between queries and utterances (Basque), conversational type of speech (Czech), and query and utterance mismatch on the level of dialect (NNEnglish). The non-native English is definitely the toughest condition.

Also very interesting is the observation of AKWS superiority for Basque and NNEnglish subsets. The reason for Basque is that as the queries are well dictated, the phoneme transcript is accurate and the AKWS provides higher accuracy, opposite to DTW where the query/utterance mismatch can be considered as a significant problem. In NNEnglish, the query conversion into phoneme string followed by search of particular sequence can be also considered as more robust for different dialects in query/utterance combinations.

It is also worth to note that the performance of AKWSDTW fusion is worse than the best of the DTW or AKWS systems for African languages, Basque and NNEnglish. This is probably due to the fact that: 1) Each African language has only $1/20$ of data and 2) the fusion is trained to maximize cross entropy. So it preferes to maximize the performance on "easy" languages with large proportions of data rather than "hard" ones having small fraction of data.

9. CONCLUSIONS

We have performed a comparison of AKWS and DTW approaches with several phone-posterior generators for QbE in several languages. We found the proposed m-norm a really promising way of score normalization of QbE systems. It seems robust and it is easy to perform in comparison to calibration based on side information and binary logistic regression. Our second interesting conclusion concerns the per-language results. It looks like DTW is well performing for high-quality speech and matching acoustic condition between query and utterance. On contrary, in case of mismatching acoustic conditions, the AKWS technique is a clear winner.

10. REFERENCES

[1] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodriguez-Fuentes, "The Spoken Web Search task," in *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[2] Florian Metze, Xavier Anguera, Etienne Barnard, Marelie Davel, and Guillaume Gravier, "The Spoken Web Search task at MediaEval 2012," in *Proceedings of ICASSP 2013*, 2013, pp. 8121–8125.

[3] Igor Szöke, Michal Fapšo, and Karel Veselý, "BUT2012 approaches for Spoken Web Search - MediaEval 2012," in *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012, CEUR Workshop Proceedings, Vol. 2012, No. 927, DE.

[4] Javier Tejedor, Michal Fapšo, Igor Szöke, Jan Černocký, and František Grézl, "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection," *ACM Transactions on Information Systems (TOIS)*, vol. 2012, no. 30, pp. 1–34, 2012.

[5] Igor Szöke, *Hybrid word-subword spoken term detection*, Ph.D. thesis, 2010.

[6] Luis J. Rodriguez-Fuentes and Mikel Penagarikano, "MediaEval 2013 Spoken Web Search task: System performance measures," Tech. Rep., Dept. Electricity and Electronics, University of the Basque Country, 2013.

[7] Petr Schwarz, Pavel Matějka, and Jan Černocký, "Towards lower error rates in phoneme recognition," in *Proceedings of 7th International Conference Text,Speech and Dialoque 2004*. 2004, p. 8, Springer Verlag.

[8] Martin Karafiát, František Grézl, Mirko Hannemann, Karel Veselý, and Jan Černocký, "BUT BABEL system for spontaneous Cantonese," in *Proceedings of Interspeech 2013*. 2013, number 8, pp. 2589–2593, International Speech Communication Association.

[9] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proceedings of IEEE 2012 Workshop on Spoken Language Technology*. 2012, pp. 336–341, IEEE Signal Processing Society.

[10] František Grézl and Martin Karafiát, "Hierarchical neural net architectures for feature extraction in ASR," in *Proceedings of INTERSPEECH 2010*. 2010, vol. 2010, pp. 1201–1204, International Speech Communication Association.

[11] Igor Szöke, Petr Schwarz, Lukáš Burget, Martin Karafiát, Pavel Matějka, and Jan Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," *Lecture Notes in Computer Science*, vol. 2005, no. 3658, pp. 8, 2005.

[12] M. Muller, *Information Retrieval for Music and Motion*, Springer-Verlag, 2007.

[13] Haipeng Wang and Tan Lee, "CUHK system for the Spoken Web Search task at MediaEval 2012," in *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012, CEUR Workshop Proceedings, Vol. 2012, No. 927, DE.

[14] Murat Akbacak, Lukáš Burget, Wan Weng, and Julien Houtvan, "Rich system combination for keyword spotting in noisy and acoustically heterogenous audio streams," in *Proceedings of ICASSP 2013*. 2013, pp. 8267–8271, IEEE Signal Processing Society.